U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

PhD

PhD

PhD

PhD

PhD

3.º
CICLO

FCUP
UNICAL
2017

U.PORTO

New Therapeutic Strategies to Lower Blood Stream Cholesterol Levels through the Inhibition of HMG-CoA Reductase

Diana Sofia Gesto da Silva

FC

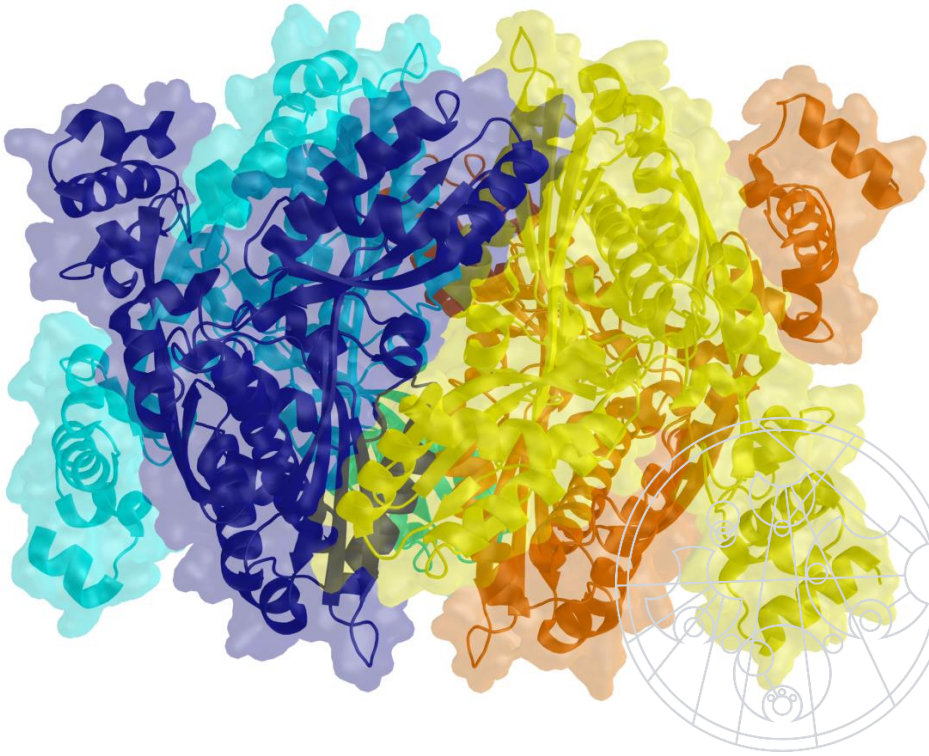
U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

UNIVERSITÀ
DELLA CALABRIA

New Therapeutic Strategies to Lower Blood Stream Cholesterol Levels through the Inhibition of HMG-CoA Reductase

Diana Sofia Gesto da Silva

Tese de Doutoramento apresentada à
Faculdade de Ciências da Universidade do Porto
Università della Calabria
Química
2017



New Therapeutic Strategies to Lower Blood Stream Cholesterol Levels through the Inhibition of HMG-CoA Reductase

Diana Sofia Gesto da Silva

Programa Doutoral em Química

na especialidade de

Química Teórica e Modelação Molecular

Departamento de Química e Bioquímica

2017

Supervisor

Pedro Alexandrino Fernandes, Associated Professor, FCUP

Co-supervisor

Nino Russo, Full Professor, UNICAL

ACKNOWLEDGMENTS

The part of the thesis everyone writes last thinking no one is going to read (and they are probably right) and that, as far as I know, one can write something either long or short, something akin to an Academy Award acceptance speech, and it will never be wrong. Still, it is something that must be written, because even though I was the one that did most of the work, the truth is it would never exist if it was not for other people as well. In that regard, I would like to leave here the names of some of those people (many more will I not be able to mention, but know that I thank you all too).

First and foremost, I would like to mention my awesome supervisor, Prof. Pedro Alexandrino Fernandes, for all the knowledge he shared with me, all the time spent at my side supporting me, and all the patient he showed towards me. He always made me feel welcome in the group and always had a kind word to say to me even on the hardest of times.

To Prof. Maria João Ramos, a well-deserved thank you, for all the guidance, leadership and sympathy she showed me over the course of these years.

To Prof. Nino Russo, of the University of Calabria, I would like to thank for accepting me in his group and helping me during my time in Italy.

To my colleagues and members of the group: I appreciate all your support and the help you gave me. I cannot name you all (it's a rather large group…) but I would like to give a special thank you to Nuno Cerqueira for helping me give my first steps in Theoretical Chemistry; to Rui Neves, Fabiola Medina, João Coimbra, Cátia Moreira and Rita Calixto, I thank you for you amazing friendship; to Gaspar Pinto, a special shout-out for putting up with me in Italy (and for all the pizzas we had together).

To all the friends I made in Italy (Valeria, Marta, Gloria, Paolo and so many others), thank you, and also to my flat mate, Macarena Valverde, for all the laughs we had together.

Now on a more personal note: I would like to thank my parents, my brother Tiago and grandparents. You are my family and I love you all. Without you I would not be here now and I know you support me no matter what are my choices. To all my friends, from all social groups (Poder, Hapkido JJK, my long-time friends) thank you so much for being there and doing what you do best: share in my adventures and laugh at their outcomes. And finally, to João Martins, thank you for being so much like myself ;)

ABSTRACT

Enzymes are exceptional biocatalysts capable of performing a wide range of reactions in all kinds of organisms. Some of them seem to work in rather mysterious ways, but science as evolved much since the first days when they were discovered and now there are several methods, both experimental and computational, that can be employed to study such remarkable molecules.

For this manuscript, we focused mostly on enzymes that are involved in the synthesis of other molecules, either fatty acids, cholesterol or aspartate. In common they have the fact that the study of their structure, mechanism of reaction and active site can help in the development of drugs that can be used in the treatment of some disease. These studies were done using theoretical and computational methods, which, albeit being fairly recent, have some advantages over experimental methods, such as the fact that they are capable of characterizing transition states. With the development of technology to the level it is currently, so did these methods evolve and keep evolving, thus allowing us to perform more accurate calculations in a smaller amount of time.

Using computational methods, we have studied several enzymes and concluded that nowadays these methods enable us to collect a wide array of data that would take a lot longer to gather by other methods. We were able to test and explore several different things, such as what is the reaction mechanism of the enzyme L-asparaginase, what residues of the interface of HMG-CoA-R are important to the formation of the tetramer or how the KS module begins the synthesis of a fatty acid molecule.

The results obtains throughout these studies can be used in future works to develop new drugs to fight diseases like hypercholesterolemia and cancer, by either searching for different molecules to inhibit the enzyme or engineering a protein we know has some effect against that disease in order to improve it.

Keywords: Enzymatic catalysis, HMG-CoA Reductase, lipids, cholesterol, QM/MM methodology, density functional theory, molecular modeling, quantum mechanics.

RESUMO

As enzimas são biocatalizadores extraordinários capazes de realizar uma grande variedade de reacções químicas nos mais diversos organismos. Algumas parecem mesmo funcionar de forma mística, mas a ciência também foi evoluindo desde que estas foram descobertas e actualmente temos à nossa disposição variados métodos; tanto experimentais como computacionais, que pode ser utilizados para estudar estas moléculas notáveis.

Neste manuscrito focamo-nos especialmente em enzimas envolvidas na síntese de outras moléculas: ácidos gordos, colesterol ou aspartato. O ponto comum entre elas é que o estudo da sua estrutura, mecanismo de reacção e centro activo pode ajudar no desenvolvimento de novos medicamentos passíveis de ser utilizados no tratamento de algumas doenças. Estes estudos foram feitos recorrendo a métodos teóricos e computacionais, os quais, apesar de serem significativamente recentes, apresentam algumas vantagens comparativamente aos métodos experimentais, tais como o facto de serem capazes de caracterizar estados de transição. Com o desenvolvimento tecnológico que se tem verificado actualmente, estes métodos também vieram a evoluir, o que nos permite realizar cálculos cada vez mais precisos num espaço de tempo cada vez menor.

Usando métodos computacionais, foi-nos possível estudar diversas enzimas e concluir que actualmente estes métodos permitem-nos chegar a uma variedade de informações diferentes, as quais iriam levar muito mais tempo a recolher com outros métodos. Foi-nos, desta forma, possível testar e explorar diversas coisas diferentes, tais como o mecanismo de reacção da L-asparaginase, quais os resíduos da interface da enzima HMG-CoA-R são fundamentais para a formação do tetrâmero ou de que forma o módulo KS inicia a síntese de uma molécula de ácido gordo.

Os resultados obtidos através destes estudos pode vir a ser usados no futuro para desenvolver novos medicamentos que nos permitam combater doenças como a hipercolesterolemia ou o cancro, usando estratégias como a procura de moléculas diferente capazes de inibir uma determinada enzima, ou melhorando uma enzima que já tenha actividade contra essa doença de forma a torna-la mais eficaz.

Palavras-chave: Catálise enzimática, HMG-CoA Reductase, lípidos, colesterol, metodologia QM/MM, teoria do funcional da densidade, modelação molecular, mecânica quântica.

LIST OF ABBREVIATIONS

ACAT	Acetyl-Coenzyme A Acetyltransferase
ACP	Acyl Carrier Protein
AMBER	Assisted Model Building with Energy Refinement
ATP	Adenosine Triphosphate
AU-ROC	Area Under the ROC curve
CADD	Computational Aided Drug Design
cASM	Computational Alanine Scanning Mutagenesis
CC	Couple Cluster
CHARMM	Chemistry at Harvard Macromolecular Mechanics
CI	Configuration Interaction
CoA	Coenzyme A
crtM	Dehydrosqualene Synthase
DFT	Density Functional Theory
DH	β-Hydroxyacyl Dehydratase
DMAPP	Dimethylallyl Pyrophosphate
DUD	Directory of Useful Decoys
EcA	E. coli Asparaginase
EF	Enrichment Factor
ER	Enoyl Reductase
FAD	Flavin Adenine Dinucleotide
FAS	Fatty Acid Synthase
FDPS	Farnesyl Pyrophosphate Synthase
FEP	Free Energy Perturbation
FPP	Farnesyl Diphosphate
FPR	False Positive Rate
GAFF	General AMBER Force Field
GPP	Geranyl Diphosphate
GTO	Gaussian-type Orbital
HDL	High-density Lipoprotein

HF	Hartree-Fock
HMG-CoA	3-hydroxy-3-methylglutaryl-coenzyme A
HMG-CoA-R	HMG-CoA Reductase
HMG-CoA-S	HMG-CoA Synthase
IHD	Ischemic Heart Disease
IPP	Isopentoyl Disphosphate
Kd	Dissociation constant
Ki	Inhibition constant
KR	β-Ketoacyl Reductase
KS	β-Ketoacyl Synthase
LDL	Low-density Protein
MAT	Malonyl/acetyltransferase
MC	Monte Carlo
MDD	Mevalonate 5-diphosphate Decarboxylase
MK	Mevalonate Kinase
MM-PBSA	Molecular Mechanics Poisson–Boltzmann Surface Area
MO	Molecular Orbital
MVAPP	Mevalonate 5-phosphate
NADP	Nicotinamide Adenine Dinucleotide Phosphate
NMR	Nuclear Magnetic Resonance
NPT	Isobaric-isothermal ensemble
NVE	Microcanonical ensemble
NVT	Canonical ensemble
ONIOM	Our own N-layered Integrated Orbital and Molecular mechanics
OS	2,3-oxidosqualene
OSC	2,3-oxidosqualene Cyclase-lanosterol Synthase
PDB	Protein Data Bank
PES	Potential Energy Surface
PLP	Piecewise Linear Potential
PMF	Potential of Mean Force

PMK	Phosphomevalonate kinase
PPT	Phosphopantetheine
QM/MM	Quantum Mechanics/Molecular Mechanics
RESP	Restrained ElectroStatic Potential
RMSd	Root Mean Square Deviation
ROC	Receiver Operator Characteristic
SASA	Solvent Accessible Surface Area
SCAP	SREBP Cleavage Activating Arotein
SCF	Self-consistent Field
SM	Squalene Monooxygenase
SMoG	SMall Molecule Growth
SQS	Squalene Synthase
SREBP-2	Sterol Regulatory Element-binding Protein 2
STO	Slater-type Orbitals
TE	Thioesterase
TI	Thermodynamic Integration
TPR	True Positive Rate
TS	Transition State
VMD	Visual Molecular Dynamics
VS	Virtual Screening
WHO	World Health Organization
ΨME	Pseudo-methyltransferase

INDEX

INDEX OF FIGURES

INDEX OF TABLES

CHAPTER 1

FATS – THE GOOD, THE BAD AND THE UGLY

1.1. The Biochemical Importance of lipids

Life as we know exists possibly due to a lot of independent coincidences that we cannot currently explain. It assumes many forms, some as simple as a single cell and other so complex that, even now, after so much scientific progress, we can still look and wonder how such amazing beings and structures came into existence. In spite of all the different forms of life that we know about today, we can still find something common in all of them, like the fact that all living organism must have some kind of genetic material, be it in the form of DNA or RNA, and a lipid membrane separating the inside environment from the outside.

All living organisms known today are carbon based, which means that the molecules that make up most of these beings have a carbon backbone. Most organic molecules essential for the correct functioning of a cell are composed of carbon, including protein, lipids, carbohydrates and nucleic acids.

Proteins can be compared to small organic machines capable of preforming a wide array of tasks in the cell. They can be quite different from one another, and yet they are all composed of the same blocks: amino acids. These small molecules bind together to form bigger molecules and depending on the amino acid content of a protein, their forms and function also change. Protein function spans from simple tasks, such as allowing certain molecules or atoms to cross the cell membrane, to other much more complex, like producing energy or carrying organelles or vesicles inside the cell. Protein can also be broken down to provide energy, but this occurs only as a last resort, since cells typically have other methods of obtaining energy.

Carbohydrates are thus called because they are made up of carbon, oxygen and hydrogen, and the ratio of oxygen to hydrogen is usually 1:2, like in water. In the cells, the major function of carbohydrates is to provide energy. They are the preferred energy source for many organisms, such as bacteria, and can either be used instantaneously or, in the case of mammals for example, stored in the form of glycogen. In humans, glucose is the primary source of energy used by the brain, and other indigestible carbohydrates, also called dietary fiber, help to regulate the gastrointestinal tract.

Nucleic acids are also some of life's most important molecules. They include both DNA and RNA, and without them it would be impossible for organisms to survive. Nucleic

acids are composed of smaller units called nucleotides, which can be assembled into huge macromolecules and carry all the information for protein assembly and the genetic instructions used in the development, functioning and reproduction of all known living organisms. RNA is a less stable molecules that is used as an intermediate between the DNA and the protein assembling mechanisms. Even though both DNA and RNA are composed of only four different nucleotides (cytosine, guanine and adenine and thymine in the DNA, whereas in RNA thymine is replaced by uracil), the different arrangements of this nucleotides in groups of three, called codons, forms the entire genetic code.

Last, but not the least, we have the lipids, a much wider and diverse class of life molecules. Lipids include molecules such as fats, sterols, fat-soluble vitamins, mono-, di- and triglycerides, amongst many others. One of the few thigs all of them have in common is that they are either hydrophobic or amphiphilic. The amphiphilic nature of some lipids allows them to form vesicles that can be used to store and transport other cellular components. This dual nature enables them also to form the cell membranes that are so important for life sustenance, as they enclose the inside of the cell and separate the cytoplasm from the outside medium. Lipids can perform a myriad of other functions in the cells. They are used to store energy, anchor proteins to the membrane, help the folding of other membrane proteins, function as co-factors, among others.

Table 1- Different types of lipids, organized in different categories according to their chemical structure, with some of their functions and examples.

Lipid Category	Functions	Examples
Fatty Acids	Energy storage	Oleate, stearoyl-CoA
Glycerolipids	Energy storage	Di- and triacylglycerols
Glycerophospholipids	Cell membrane building blocks	Phosphatidylcholine, phosphatidylserine
Sphingolipds	Cell protection	Sphingomyelin
Sterol Lipids	Varied (ex: cell membrane structure, hormones)	Cholesterol, bile acids
Prenol Lipids	Varied (ex: antioxidants agents)	Farnesol, geraniol
Saccharolipids	Cell membrane structure	Lipopolysaccharide
Polyketides	Varied (ex: antibiotic)	Tetracycline, aflatoxin B_1

Lipids can be divided in different categories according to their chemical structure (Table 1). Having into account that lipids can be very different from one another, they can also be grouped according to their function. In the following sections, we will further discuss their most important functions in the cells and in mammals.

1.2. Fatty acids

Maybe their best known function, being used as cell fuel is a very important task that lipids perform in many organisms. Fatty acids are the basic precursors of most lipids. Contrarily to proteins and nucleic acids, lipids do not have a monomers that come together in order to form large macromolecules. Instead, they start as a smaller lipid, usually fatty acids, that are joined to others but in small number (mostly 2 or 3 at a time) by a glycerol molecule. This gives rise to different classes of lipids, such as glycerophospholipids or triglycerides. Fatty acids can also be metabolized into other lipids.[1]

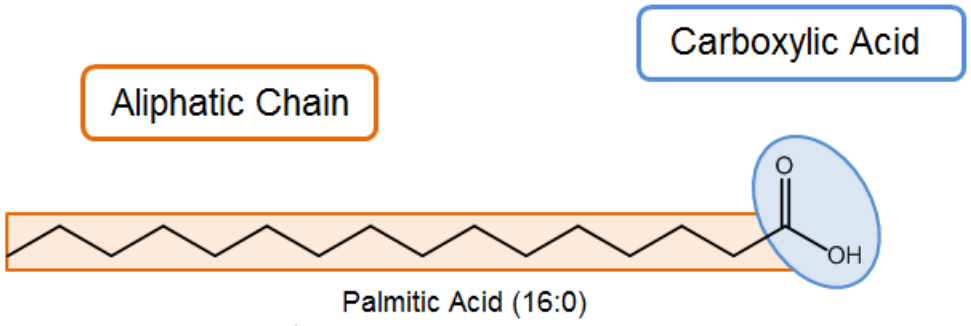


Figure 1 – Structure of the palmitic acid as an example of the dual nature of fatty acids. In orange is represented the large aliphatic (hydrophobic) chain and in blue the carboxylic head group (hydrophilic).

Fatty acids are composed of a long hydrocarbon chain attached to a carboxyl group. This causes the fatty acids to be amphiphilic in nature, having a hydrophilic head group and a lipophilic tail end (Figure 1). Their length can vary within the cells, but the most common fatty acids in the cells have either 14, 16, 18 or 20 carbon atoms. For storing energy, three fatty acid molecules are combined with glycerol, they form triacylglycerides, which are stored as fat droplets in the adipose cells. Whenever the organisms are in need of energy, and in response to hormones such as adrenaline, triacylglycerides are hydrolyzed in the cytosol and release the fatty acids into the blood. These, in turn are taken up by the cells and oxidized to CO_2 to yield energy.

Even though the hydrophilic head of fatty acids is always the same, a carboxylic group, their lipophilic end can change in both size and saturation. Some other less conventional fatty acids can even have branched chains or contain carbon rings. Most naturally occurring fatty acids have a length that ranges from 12 to 24 carbons in a linear configuration. Double bounds are also frequent in some structures, and usually occur between the ninth and tenth carbon in monounsaturated fatty acids, and in the case of polyunsaturated fatty acids the other double bounds usually form between C-12 and C-13, and C-15 and C-14. Double bounds rarely occur conjugated in polyunsaturated fatty acids (alternating between single and double bonds, as in –CH=CH–CH=CH–) but are frequently separated by a methylene group (as in –CH=CH–CH_2–CH=CH–). Other common characteristic present in most naturally occurring unsaturated fatty acids is the fact that the double bonds are usually in the cis configuration. Trans fatty acids are a product of fermentation in the rumen of certain animals, and can be obtained in dairy products.

The physical properties of fatty acids depend mostly on both their length and degree of saturation. [1] For example, solubility is dependent on the number of carbons in the structure; the larger the number, the less soluble the fatty acid is. This is easily understandable, since the only part of the fatty acid that is hydrophilic is the carboxylic group. If the chain is small enough, the fatty acid can have some solubility in water. However, since the carbon chain is completely hydrophobic, the larger it is, the smaller its solubility.

Other property that varies greatly amongst fatty acids is their melting point. This property is not so much related to the length of the carbon chain but rather to its saturation degree. It is evident that the larger the fatty acid the higher its melting point will be, since larger molecules have a bigger contact area and therefore the number of Van der Waals interactions will also be bigger. On the other hand, in the case of saturated versus unsaturated fatty acids, if both have the same number of carbons, the saturated one will have the higher melting point. This too relates to the Van der Waals energy, but in this case it has to do with their ability to be packed. Saturated fatty acid are much more flexible and can therefore be packed much more neatly and tightly, forming almost crystalline arrangements (Figure 2a). Unsaturated fatty acids, contrarily, have a more rigid structure, resulting from the bend imposed by the double bond (Figure 2b). This results in their inability to form connections as tight as their saturated counterpart, meaning that the energy needed to melt them will be smaller.

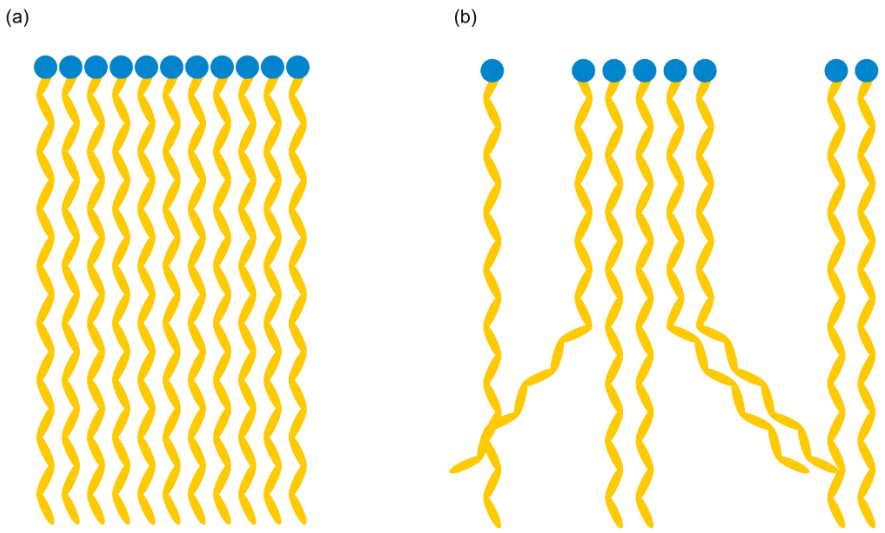


Figure 2 – Representation of how the saturation level of fatty acids affects the way they are packed. (a) Saturated fatty acids can be packed more tightly and neatly, whereas when unsaturated fatty acids are also present (b) the bend in their structure makes it more difficult for them to organize in a neat structure.

1.2.1. Fatty Acids as an Energy Source

Fatty acids constitute the major source of energy in many tissues, and especially in the heart muscle. In humans, the oxidation of fatty acids is quantitatively more important than that of glucose, as a source of ATP. The reason for this is that fats are stored in a much more reduced form that carbohydrates.

The conversion of fatty acids into energy (ATP) is called β-oxidation and occurs in the mitochondria[2-3]. A similar process occurs in the peroxisomes, although in this case it is not coupled to ATP production and is used as a means of shortening very long fatty acids (20+ carbon long) before these can be relocated to the mitochondria where β-oxidation proceeds as usual (Figure 3). Since peroxisomes lack respiratory chain, contrarily to mitochondria, the energy produced during oxidation is dissipated as heat. Despite this major difference, both oxidative processes are basically identical, and are mediated by similar enzymes. They begin with the esterification of the fatty acid molecule to coenzyme A (CoA) in the cytosol, forming fatty acyl CoA. This initial reaction requires energy, meaning that a molecules of ATP is hydrolyzed to AMP. Next, the fatty acyl CoA is relocated to either the mitochondria or the peroxisome, where it undergoes several cycles of four reaction each. During each cycle the fatty acyl CoA is shortened by two carbons, which are converted into acyl CoA and the remaining fatty acyl CoA, with the generation of NADH and $FADH_2$.

This process is repeated sequentially until all the fatty acid has been used. The complete oxidation of an 18-carbon fatty acid yields eight molecules of NADH and eight of $FADH_2$ as well as nine molecules of acyl-CoA. The acyl-CoA produced during β-oxidation then undergo the citric acid cycle and are oxidized to CO_2. The $FADH_2$ and NADH produced during both the citric acid cycle and the β-oxidation of fatty acids are used to create a proton-motive force, which in turn powers the synthesis of ATP. It is through this whole process that the cells create energy through lipids.

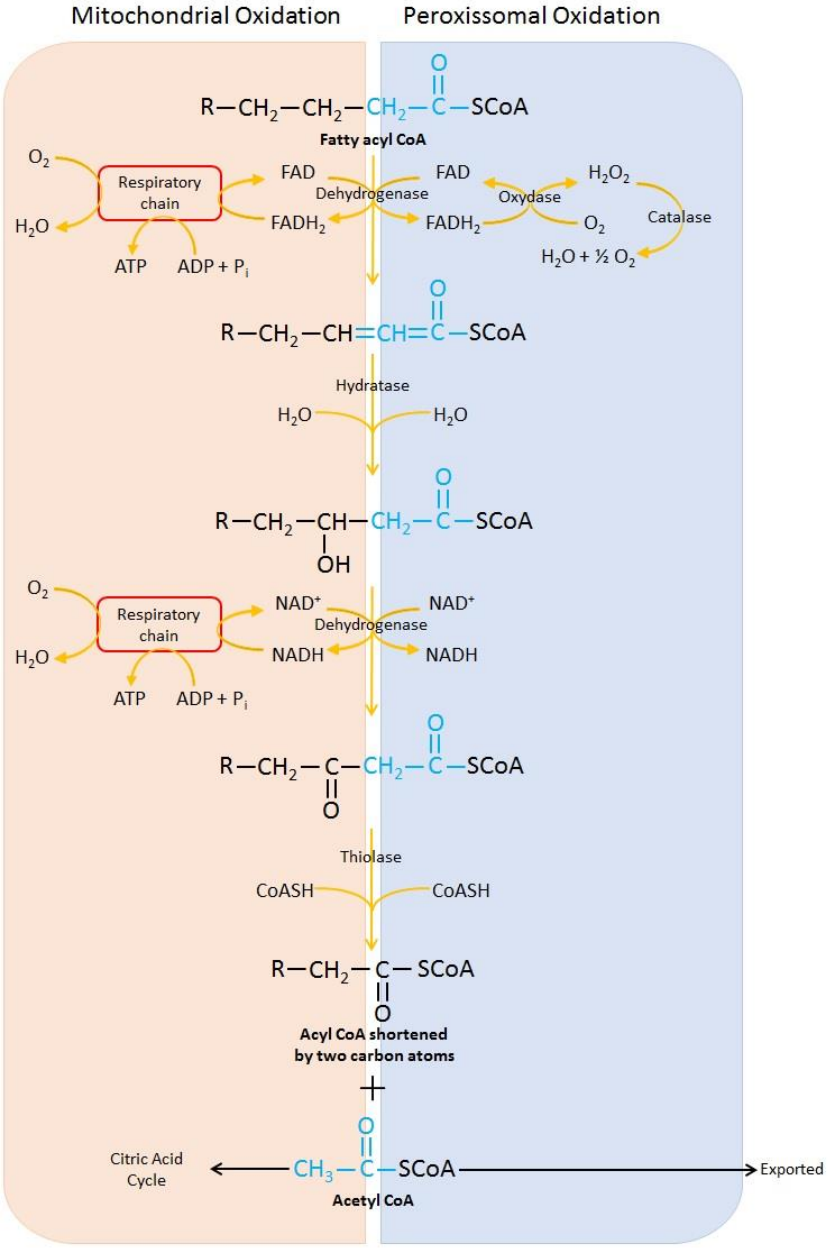


Figure 3 – Schematic representation of the oxidation of fatty acids in both mitochondria and peroxisomes

1.2.2. Fatty Acids as Precursors to other Lipids

1.2.2.1. Triacylglycerides

As stated previously, lipids can take up many forms and play many different roles in the organism. Fatty acids are probably the most versatile of lipids, as they can be easily modified and transformed into different kinds of lipids or some other molecule.

Fatty acids can bind to a molecule of glycerol to form triglycerides (Figure 4). The bond formed is an ester linkage in which three fatty acids are connected to a single glycerol. Triglycerides can contain only one type of fatty acid, in which case they are referred to as simple triglycerides, or they can be composed of a mix of different fatty acids.

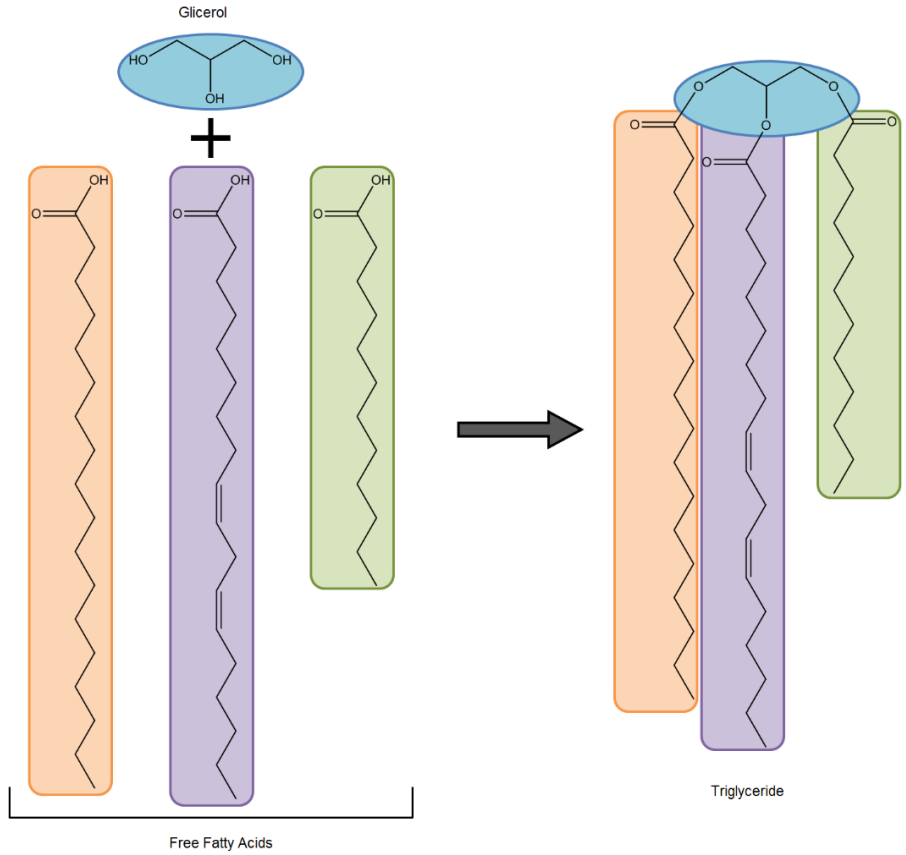


Figure 4 – Simplified representation of how a triglyceride is formed from three free fatty acid and a glycerol molecule.

Triglycerides are mostly non-polar, because the polar carboxylates from fatty acids and the polar hydroxyls from glycerol are both bond in the ester linkage. This renders

the triacylglycerides practically insoluble in water. In mammals and other vertebrates, the triglycerides are stored in special cells called adipocytes, which in turn form adipose tissue. Since they are not soluble in the aqueous cytoplasm, stored fats form big droplets that occupy most of the adipocyte cell.

Most organisms use fats as a form of stored energy because they present several advantages in comparison to polysaccharides: they are dehydrated, which means they can be stored without the organism having to carry around the extra weight of the water, which is needed in the case of polysaccharides, and also the oxidation of one gram of fat yields more than the double of the energy produced by the oxidation of one gram of carbohydrates[1]. For this reasons triacylglycerides are the primary molecule for storing energy, whereas carbohydrates are mostly used as a quick source of metabolic energy. For animals living in cold climates, these stored fats have the additional function of insulator, preventing the body from losing too much heat.

1.2.2.2. Phospholipids and Glycolipids

Another type of lipids that can be obtain from fatty acids are phospholipids and glycolipids. These types of lipids are vital for all organisms, as they are the main constituents that compose cell membranes. Although different in structure, both phospholipids and glycolipids have similar characteristics, and they both also contain one or more types of fatty acids in their structure.

Cell membranes are very complex systems that separate the cells cytoplasm from the external medium, while still allowing for the transference of compounds from the inside to the outside and vice-versa, as well as carrying out other important functions. They comprise some very different components, although, in a simplistic way, they can be seen as composed essentially of a double layer of phospholipids.

Membrane lipids are amphiphilic, which means that one of their ends is polar and hydrophilic and the other one is non-polar and lipophilic (Figure 5). It is this dual nature of membrane lipids that gives biological membranes their characteristic architecture, the double layer of lipids. Being amphiphilic, membrane lipids tend to interact in a way so that their lipophilic ends interact with each other and are oriented to the middle of the membrane, whereas the hydrophilic ends interact with the water in the environment or the aqueous cytoplasm. Membranes act as a barrier to prevent the passage of polar molecules from one side of the membrane to the other, and this architecture is indeed very efficient at accomplishing this task.

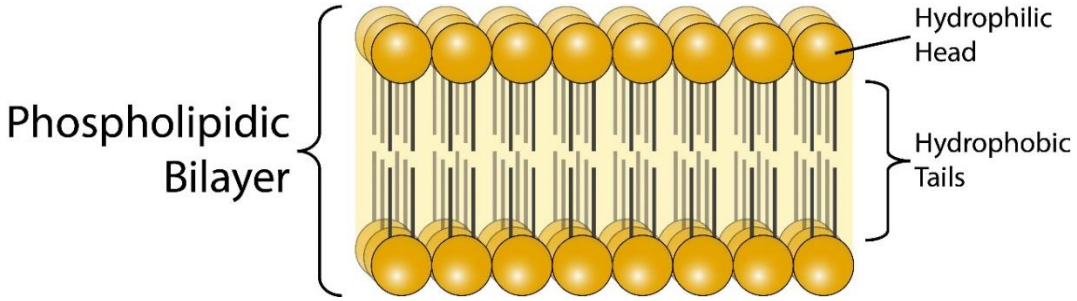


Figure 5 – Illustration of how phospholipids are arranged to form a phospholipid bilayer. Their head groups (polar) are oriented towards the outside so they can interact with the medium, while their tails (apolar) are on the inside and interact with each other to reduce unfavorable interactions.

Membranes from most organisms are composed primarily of two different types of lipids: phospholipid and glycolipids (Figure 6). In phospholipids the polar head group is connected to the non-polar tail by a phosphodiester linkage. On the other hand, the hydrophobic head is composed of a simple sugar or a complex oligosaccharide.

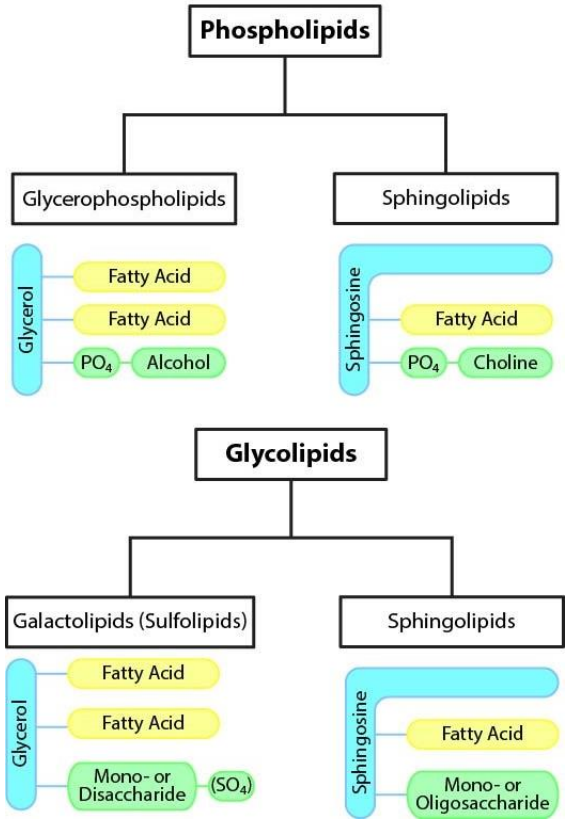


Figure 6 – Division of the different types of lipids that mainly compose the biological membranes of most organisms.

Phospholipids can be further divided in two different classes: glycerophospholipds, if they are derived from a glycerol molecule, or sphingolipids, if they are derived from sphingosine. Glycerophospholipids have two fatty acids attached to the first and second carbon of the glycerol molecule through an ester linkage. This constitutes the non-polar part of the phospholipid. The polar head is bound to the last carbon from glycerol and it comprises a highly polar or charged alcohol (some examples are choline and ethanolamine) linked via a phosphodiester linkage. At neutral pH, the hydrophilic head group can be either neutral, positively or negatively charged, which contributes to some of the properties of the cellular membrane. The fatty acids that constitute the lipophilic tail can vary significantly and can sometimes be specific for different organisms or tissues, but in general the C1 from glycerol contains either a C_{16} or C_{18} saturated fatty acid and the C2 contains a C_{18} or C_{20} unsaturated fatty acid.

Contrarily to glycerophospholipds, sphingolipids do not derive from glycerol but rather from sphingosine (Figure 7). Sphingosine is an amino alcohol that contains a large aliphatic chain. Carbons C1, C2 and C3 of sphingosine are structurally similar to those of glycerol. The long aliphatic chain is attached to C3 and play the same role a fatty acid would play in a glycerophospholipid. C2 is connected to a nitrogen which is able to bond to a fatty acid through an amide linkage. Finally, C1 can bind different compounds. In the case of phospholipids, C1 binds choline through a phosphodiester linkage. The choline then becomes the polar head group of the phospholipid, whereas the aliphatic chain of sphingosine and the fatty acid constitute the hydrophobic tail group.

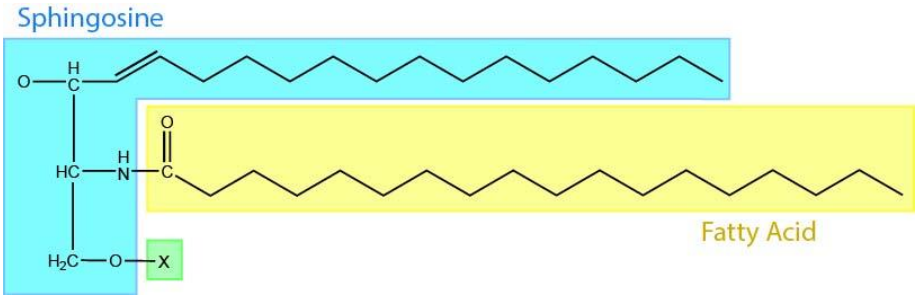


Figure 7 – Sphingolipid structure: the blue part represents the sphingosine molecule, the yellow is a fatty acid that binds the nitrogen atom of the sphingosine and the X (green) can be different modifications (either a hydroxy or a variety of carbohydrate or phospholipid structures)

Instead of choline, C1 can also bind one or more sugars directly to the –OH. This compounds are called glycosphingolipids and occur largely on the outer face on the plasma membrane. Since glycosphingolipids do not contain phosphate but they do comprise a hydrocarbon molecule they belong to the glycolipids group.

A different type of glycolipids is the galactolipids. These are predominant in plant cells and have one or two galactose residues connected to the C3 of a glycerol molecule by a glycoside linkage. Galactolipids are found mostly on the inner membrane of the chloroplast.

1.2.3. Fatty Acid Synthesis

While we can obtain a many fatty acids through the food we eat, our body and that of many organisms is also capable of synthetizing them from smaller molecules. Fatty acid synthesis is a rather complex process that requires various steps that occur cyclically. Even though the chemistry behind fatty acid synthesis is more or less conserved in all organisms, the enzymes that catalyze them can be somewhat different. For example, fatty acid synthesis in bacteria is catalyzed by several different enzymes, whereas in mammals it is accomplished by a single enzyme called fatty acid synthase (FAS)[4].

The process of fatty acid synthesis always begins with the same two molecules, acetyl and malonyl. Only one acetyl molecule is needed for the synthesis of one fatty acid, while the number of malonyl varies according to the final number of carbons in the fatty acid. Both acetyl and malonyl used in the biosynthesis of fatty acids derive from acetyl-CoA and malonyl-CoA respectively. NADPH is also needed to function as a reducing agent during this process. The complete reaction catalyzed by FAS is summarized in the following equation:

$$\text{Acetyl-CoA} + 7\ \text{Malonyl-CoA} + 14\ \text{NADPH} + 14\ \text{H} \longrightarrow \text{Palmitic Acid} + 7\ \text{CO2} + 8\ \text{CoA} + 14\ \text{NADP}^{+} + 6\ \text{H2O}$$

The mammalian FAS and the reaction catalyzed by it will be further discussed in chapter 7.

1.3. Cholesterol: a Vital Molecule

The cholesterol molecule is of vital importance to almost all forms of life known to date, especially for animals. However, due to its association with several heart diseases, cholesterol is currently perceived as a somewhat bad compound that needs to be

avoided at all costs. The truth is, in fact, that without cholesterol, life has we know it would not be possible.

Cholesterol (Figure 8) is the major sterol present in animal tissues. It is amphiphilic, having both a hydrophobic hydrocarbon body and a hydrophilic head group which includes a hydroxyl group. Since it belongs to the sterol group, cholesterol is a structural lipid and can be found in the cellular membrane of most eukaryotic cells. It contains a steroid nucleus that consists of four fused rings, three of which have six carbons and one with five, a feature shared between all sterols. This nucleus is also characterized for being almost planar and rigid.

In mammals, cholesterol plays a vital role and is considered an essential component for the proper functioning of our cells. Its roles range from component in cell membranes to precursor of several steroid hormones[1].

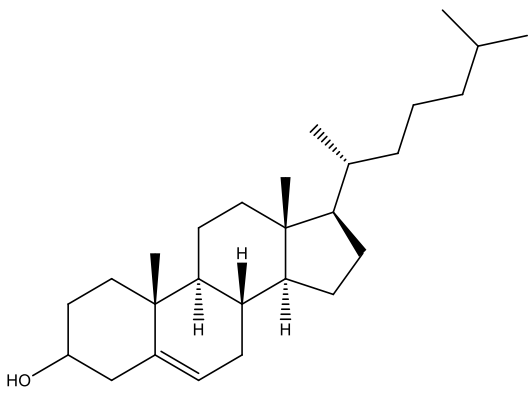


Figure 8 – Structure of the cholesterol molecule.

The importance of cholesterol in cell membranes can be assessed by comparing the fluidity of membranes with more or less cholesterol. The polar head group of cholesterol interacts with the similar head group of the phospholipids, while the nonpolar group interacts with their hydrophobic tails. Since cholesterol is somewhat rigid, membranes with larger cholesterol content will tend to be more rigid and packed, while those with less cholesterol will be more fluid. Cholesterol is also important in other membrane processes, such as endocytosis[1].

In addition to being a structural component of membranes, cholesterol is also a precursor of natural steroid hormones produced in our body. These include hormones produced in the adrenal gland, male sex hormones and female sex hormones. The adrenal gland is located just above the kidneys and its main role is to produce and release two classes of hormones: corticosteroids and catecholamines (epinephrine and

norepinephrine). From these only the first is derived from cholesterol and it can be further divided into two other classes: mineralocorticoids and glucocorticoids. Mineralocorticoids, like aldosterone, are hormones related to the reabsorption of such ions as Cl^-, Na^+ and HCO_3^- by the kidney, therefore regulating the concentration of electrolytes in the blood. On the other hand, glucocorticoids, like cortisol, affect above all the metabolism of carbohydrates in the body, regulating the metabolism of glucose. Furthermore, sex hormones produced by male and female sex glands are also derived from cholesterol. These can be divided into androgens, such as testosterone, which are responsible for the development of male characteristics, and estrogens and progestogens, which are responsible for the development of female characteristics among other functions[1].

Other fates of the cholesterol molecule in our bodies are as precursor in the synthesis of bile salts and also in the synthesis of vitamin D[1].

Cholesterol present in our bodies derives from two different sources: it can either be synthesized *de novo* within our cells, or it can be obtained through ingestion of certain foods, such as beef and pork meat, eggs and cheese. Although many people regularly eat these foods, there really is no absolute need to ingest them for the sole purpose of obtain more cholesterol, since our own cells are capable of producing enough quantity of this molecule for everything our body needs[5-6]. Nonetheless, whether or not there is dietary intake of cholesterol, its levels are maintained through regulation of both synthesis and absorption, which means that when low quantities of cholesterol are ingested, absorption and synthesis will be up-regulated. Likewise, if dietary intake is high, its excretion will be increased[7].

Currently, cholesterol has gained a bad reputation next to the great majority of people, especially because of its association with cardiovascular diseases. According to the World Wealth Association (www.who.int), in the top 10 leading causes of death (worldwide) in 2008, ischemic heart disease (IHD) was number one, accounting for 12.8% of total deaths, followed by stroke and other cerebrovascular disease as number two (10.8%). Since both of these diseases are associated with hypercholesterolemia, it is easy to understand why this molecule has such bad name.

In a more direct way, cholesterol is related to atherosclerosis[8], which is one of the main causes of both IHD, stroke and cerebrovascular diseases. Atherosclerosis is nothing more than the accumulation of fatty materials, such as cholesterol, on the walls of arteries. It is a complex process which involves a chronic inflammatory response on the walls of arteries to oxidized low-density proteins (LDL). LDL are very rich in

cholesterol and cholesteryl esters and, when oxidized, they are toxic to the cells on the walls of blood vessels, triggering an inflammatory response. This leads to a pathogenic accumulation of cholesterol in blood vessels and the formation of atherosclerotic plaques, resulting in the constriction of blood vessels[8]. Atherosclerosis occurs when the amount of cholesterol in the blood, due to either unregulated synthesis or considerable ingestion of cholesterol rich foods, exceed the amount need for the production of steroids, bile acids and membranes.

1.3.1. Biosynthesis of Cholesterol and the Role of HMG-CoA

The biosynthesis of cholesterol is a complex process, heavily regulated at several points throughout its progression. Some of the intermediaries can be diverted and used as precursors in the biosynthesis of other compounds or perform themselves certain functions on the body[9]. This process requires numerous enzymes, some of which are accounted amongst the most regulated enzymes known, as is the case of 3-hydroxy-3-methylglutaryl-CoA reductase (HMG-CoA-R)[10].

The first step in the synthesis of cholesterol is the formation of mevalonate from acetate (Figure 9)[1]. It begins with the condensation of two acetyl coenzyme A (acetyl-CoA) molecules to form acetoacetyl-CoA, a process catalyzed by the enzyme thiolase. Next, HMG-CoA synthase catalyzes the reaction between acetoacetyl-CoA and another molecule of acetyl-CoA in order to form HMG-CoA. The final step in the synthesis of mevalonate is accomplished by the enzyme HMG-CoA reductase and it is not only the committed step of this whole process but also the rate-limiting one. The HMG-CoA is reduced to mevalonate by the enzyme with the aid of two NADPH.

The subsequent step in the biosynthesis of cholesterol comprises the conversion of mevalonate into two activated isoprenes (isopentanyl-5-pyrophosphate and dymethylallyl pyrophosphate). Following a series of successive condensations of activated isoprenes, a 30 carbon molecule will be formed: squalene. Squalene is the biochemical precursor of all steroids, and despite being a linear compound, its structure can still be linked to that of the cyclic steroids. In order to form cholesterol, squalene has to endure a succession of changes, being initially converted to lanosterol, a 4 ring compound, which is finally transformed into cholesterol after about 20 reactions.

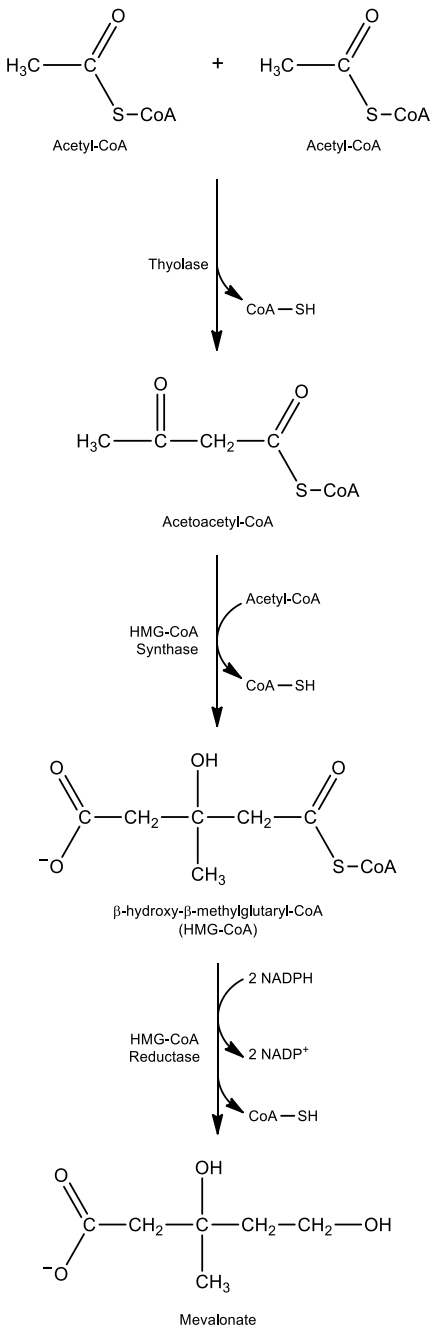


Figure 9 – Diagram of the mevalonate pathway, first step in the biosynthesis of cholesterol.

Since cholesterol biosynthesis is an energetically expensive and complex process, it is only natural to assume that it is also carefully regulated. In order for this to happen, there are several checkpoints throughout the biosynthesis of cholesterol. In mammals, its regulation is controlled by both intracellular cholesterol concentration and hormones (insulin and glucagon)[10]. Perhaps the most important of all these checkpoints is the conversion of HMG-CoA to mevalonate, a step catalyzed by HMG-CoA-R, as mention before. The importance of this enzyme to the mevalonate pathway has been evaluated through various experimental works, including that of Chappell (1995) in which, after

introducing a constitutively expressed HMG-CoA-R gene from a hamster into tobacco plants, the activity of the enzyme became unregulated and the accumulation of sterols increased 3- to 10-fold[11]. Because of this reason, and some others about to be discussed, this enzyme has been subject to many studies, made with the intention to assess not only its importance on the biosynthesis of steroids, and above all cholesterol, but also to understand how it works and is regulated.

1.3.1.1. Structure of HMG-CoA Reductase

Isoprenoids, the main product of the mevalonate pathway, are of major importance in a lot of different organisms, from bacteria to plants and animals[12]. For this reason, many of these organisms also have similar enzymes that perform analogous reactions.

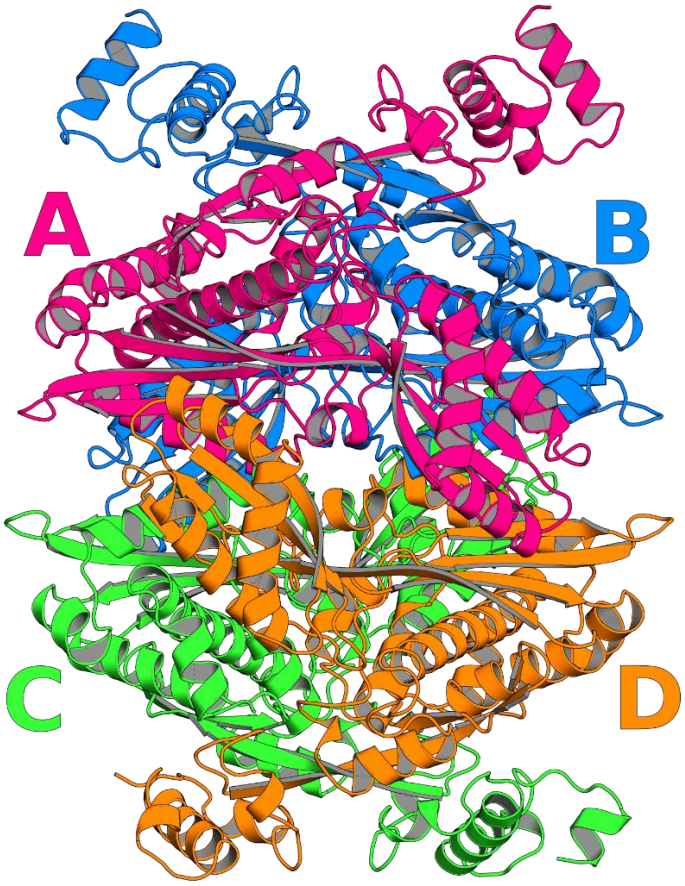


Figure 10 – Structure of the tetramer of HMG-CoA reductase.

HMG-CoA-R can be found in both eukaryotes and prokaryotes, including mammals and plants. From sequence analysis it was possible to divide different HMG-CoA

reductases in two classes. Class I enzymes comprise those from eukaryotes and the majority of archaea while class II enzymes are found in prokaryotes and some archaea[13]. Although the amino acid sequence of the catalytic portion is well conserved within each classes, the same cannot be said when classes are compared between themselves, with identities ranging between 14 to 20%[14].

In bacteria, the enzyme is soluble in the cytoplasm[15], whereas in all other eukaryotic organisms it is inserted in the membrane of the endoplasmic reticulum and has a transmembrane domain (residues 1 to 339 in humans) and a cytosolic domain (residues 460 to 888 in humans), where the active site is located, connected to each other by a linker region (residues 340 to 459). Prokaryotic HMG-CoA-R lacks both the transmembrane domain and linker.

The tridimensional structure for the catalytic domain human HMG-CoA-R (Figure 10) has been solved for the first time in 2000 by Istvan et al[16]. Currently there are over 20 structures for the human enzyme available on the Protein Data Bank website, totaling 27 if we also count those from other organisms. They include not only structures from the enzyme alone but also complexes between HMG-CoA reductase and HMG-CoA, $NADP^+$ and different statins (Table 2).

Table 2 – Current known HMG-CoA reductase crystal structures available in the Protein Data Bank. The table is ordered by increasing resolution.

PDB	Year	Resolution (Å)	Size	Ligand(s)	PDB	Year	Resolution (Å)	Size	Ligand(s)
2R4F	2008	1.70	441	RIE	3CCW	2008	2.10	441	4HI
3CCZ	2008	1.70	441	5HI	3CCT	2008	2.12	441	3HI
1DQA	2000	2.00	441	HMG, CoA, NADP	1HWK	2001	2.22	467	ADP, 117
2Q6B	2007	2.00	441	HR2	3BGL	2008	2.22	441	RID
2Q6C	2007	2.00	441	HR1	1HWJ	2001	2.26	467	ADP, 116
2Q1L	2007	2.05	441	882	1HWI	2001	2.30	467	ADP, 115
3CD7	2008	2.05	441	882	3CDB	2008	2.30	441	9HI
3CDA	2008	2.07	441	8HI	1HW9	2001	2.33	467	ADP, SIM
1DQ8	2000	2.10	467	HMG, CoA, DTT	3CD5	2008	2.39	441	7HI
1HW8	2001	2.10	467	ADP, Compactin	3CD0	2008	2.40	441	6HI
1HWL	2001	2.10	467	ADP, FBI	1DQ9	2000	2.80	441	HMG

From the structures resolved until now we can see that human HMG-CoA reductase is a tetramer produced by four identical monomers. The monomers form

dimers in which each subunit is coiled around the other in an intricate way (Figure 11). Each enzyme contains four active sites, two in each dimer, and they are made up of residues from both subunits. The monomer (Figure 12) itself can be divided in three different domains: a small, helical amino-terminal N-domain (residues 460 to 527), a large central L-domain that binds HMG-CoA (residues 528 to 590) and a small carboxyl-terminal S-domain that binds NADPH (residues 591 to 682).

The dimer interfaces are extensive and all domains of the monomer participate in the interactions that join both subunits together. However, the most broad interactions are located in the three following regions[17]:

- The loop that connects the L-domain and the S-domain (residues 682 to 694), called the *cis*-loop.
- The region in the L-domain where an intramolecular β-sheet is formed. The two strands of this β-sheet are characterized by a highly conserved sequence (ENVICX_3I/LP)
- A four α-helix bundle (two from each monomer) formed by helixes Lα6 and Lα7, which fold in an antiparallel fashion with the corresponding helixes from the neighboring subunit.

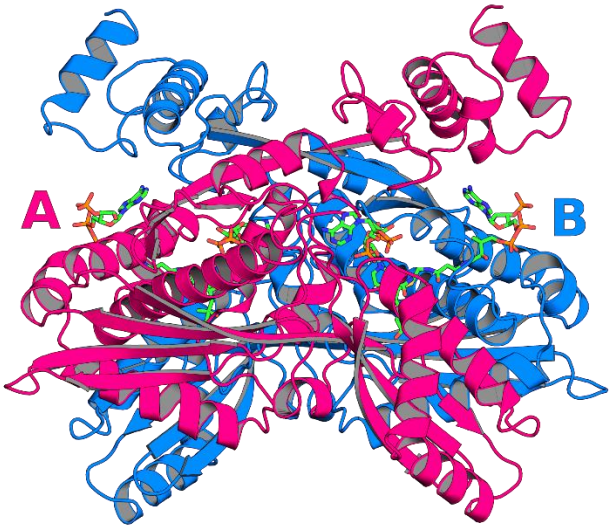


Figure 11 – Structure of one dimer of HMG-CoA reductase.

Despite the fact that the HMG-CoA reductase tetramer was only observed on crystallized structures of the catalytic domain alone, other experiments suggest that this

configuration is maintained even in the full-length human HMG-CoA reductase, containing both catalytic and membrane domains[16].

Figure 12 – Illustration of the monomer of HMG-CoA reductase, with evidence on the different domains. In red is represented the N-domain, in orange the L-domain and in yellow the S-domain. It is also possible to see the binding site for both NADPH (upper right) and HMG-CoA (lower left).

The membrane domain of HMG-CoA reductase is not very conserved in eukaryotes, contrarily to the catalytic portion. In humans, the membrane spanning region forms 8 helixes that are inserted in the membrane. Mammalian enzymes contain a 167-residue segment which is sensitive to sterols. This segment has a sequence identity of approximately 25% with other enzymes that are influenced by cholesterol[18].

1.3.1.2. Active Site Architecture and Catalytic Mechanism of HMG-CoA Reductase

The active site of HMG-CoA reductase is formed by residues of two different subunits, which form a dimer when bound together. In the same active site, HMG-CoA binds to the L-domain of one monomer, while NADPH binds to the S-domain of the other. The formation of the tetramer does not appear to be involved in the substrate binding.

Multiple interactions are formed between the L-domain of one monomer and the HMG-CoA moiety. The CoA portion of the substrate binds in an extended conformation.

The ADP portion of CoA interacts with a positively charged pocket near the enzyme surface. Residues Ser565, Asn567, Arg568, Lys722, Ser865 and His866 are involved in the binding of CoA (Figure 13A). There is, however, one residue from the neighboring subunit (Tyr479) which interacts with CoA. The side chain of Tyr479 stacks against the adenine base in van der Waals interactions while the hydroxyl group makes a hydrogen bond with the 3'-phosphate of the ribose moiety.

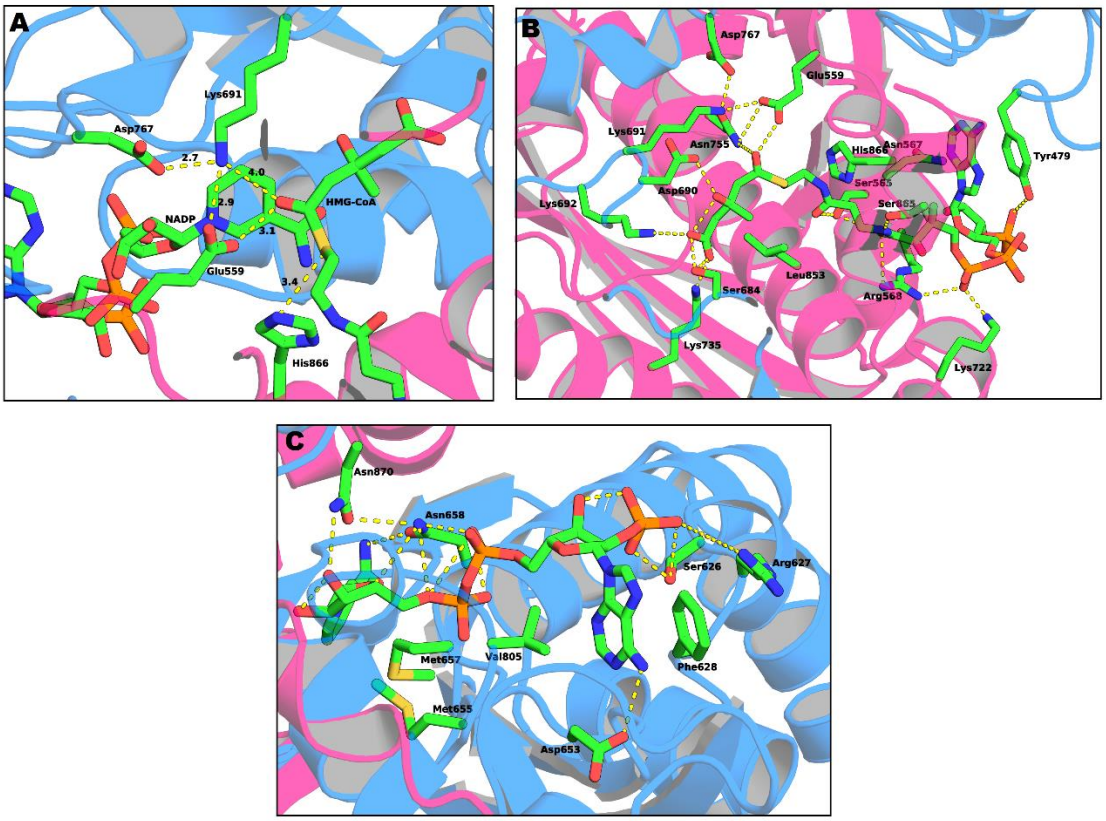


Figure 13 – Representation of the active site of HMG-CoA-R (A), the binding site for HMG-CoA (B) and NADPH (C). Each subunit of the dimer is colored differently (pink and blue).

NADPH binds essentially to the S-domain of the other subunit. Residues Ser626, Arg627, Phe628, Asp653, Met655, Gly656, Met657, Asn658, Val805 from one monomer and residues Asn870 and Arg871 from the neighboring one form the binding pocket were NADPH binds (Figure 13B). The diphosphate group is stabilized by the conserved sequence element DAMGMN.

The HMG binding pocket is located between the L- and S- domain and residues from both subunits contribute to the binding. This is also the catalytic site. HMG make a sort of a bridge between HMG-CoA and NADPH. The binding pocket of HMG is formed

by residues Ser684, Asp690, Lys691, Lys692 and Asp767 from one subunit and Glu559, Lys735, Asn755, Leu853 and His866 from the other (Figure 13C).

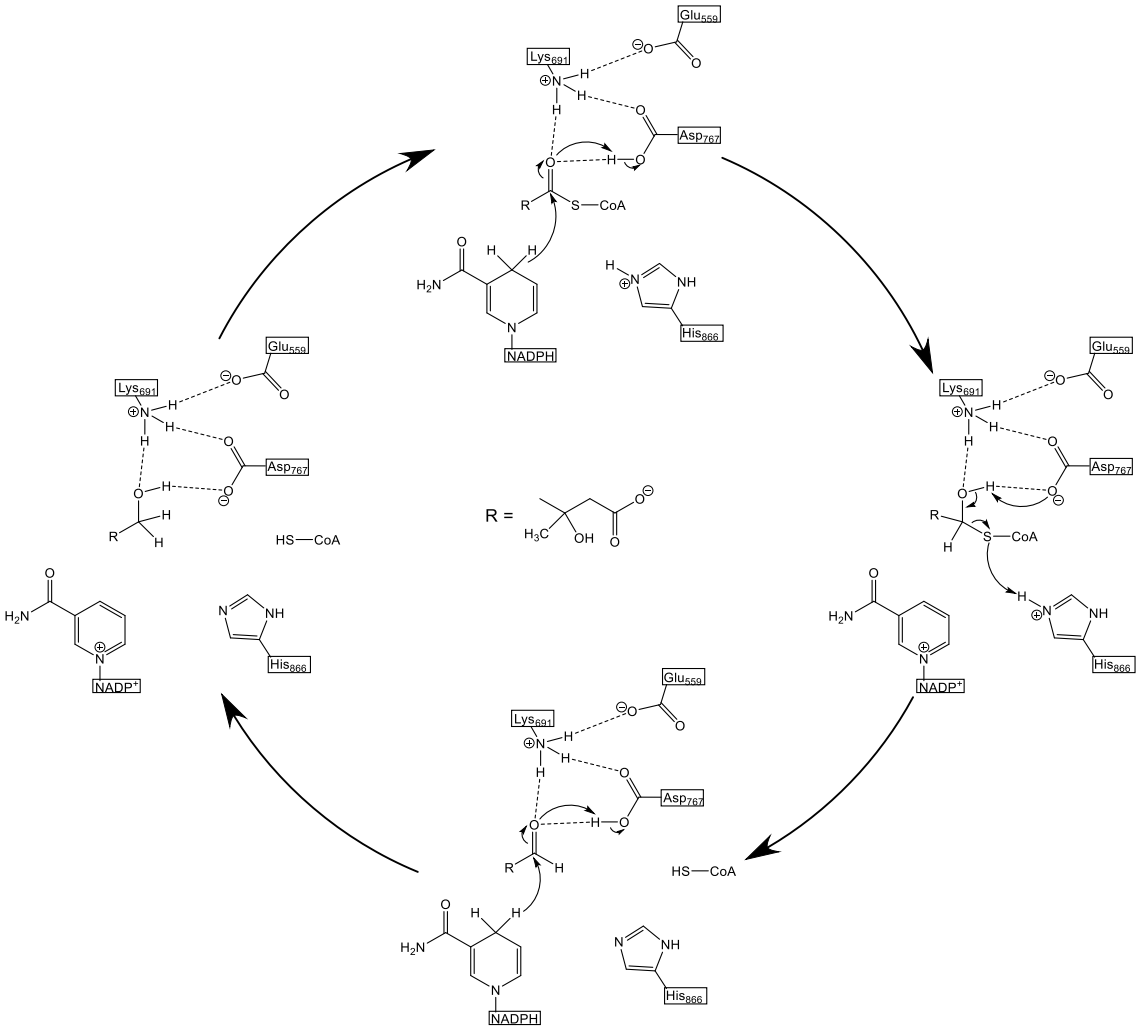


Figure 14 – Currently accepted catalytic mechanism of HMG-CoA reductase. The catalytic residues are Lys691, Glu559 and His866.

The side chains of both Lys691 and Glu559 are ideally positioned in the active site to participate directly in the reduction of HMG-CoA. Negatively charged intermediates can be stabilized by the positive charge of Lys691. The proximity between one of the side chain oxygens of Glu559 and the carbonyl oxygen of HMG may suggest that this residue is protonated. After the binding of NADPH, His866 can make a hydrogen bond with the CoA thiol, which is consistent with the role proposed for this amino acid in the currently accepted mechanism, where histidine donates a proton to the thioanion[19]. The proposed mechanism is represented in Figure 14.

1.3.1.3. Regulation of HMG-CoA Reductase

HMG-CoA reductase is one of the most regulated enzymes in our organism[10]. Its regulation can be achieved in four different ways: transcription of the enzyme's gene[20-21], translation of its mRNA[22], degradation of the functional enzyme[23] and modulation of its activity[24].

Transcription of HMG-CoA reductase gene is controlled by the transcription factor sterol regulatory element-binding protein 2 (SREBP-2)[25]. It regulates the levels of mRNA in response to sterols. Initially, SREBP-2 binds itself to the SREBP cleavage activating protein (SCAP) in the endoplasmic reticulum or in the nuclear envelope to form the complex SCAP-SREBP, which is sensitive to sterol levels. When the concentration of cholesterol in the cell is high enough, proteins Insig-1 and Insig-2 bind to the SCAP portion of the complex, preventing its movement from the reticulum to the Golgi. However, if the levels of cholesterol in the cell are low, these two proteins allow activation of the complex and its translocation to the Golgi, where the SREBP will be sliced in two positions. This cleavage releases the N-terminal basic helix-loop-helix domain, which is now free to enter the nucleus and behave as a transcriptional factor and is able to recognize certain sequences of DNA called sterol-regulatory elements. When this transcription factor binds them, it promotes the transcription of the HMG-CoA reductase gene.

The regulation of the degradation of HMG-CoA reductase involves the membrane portion of the enzyme[26] and is induced by a non-sterol metabolite derived either from mevalonate alone or mevalonate and another sterol[10]. Once again, protein Insig-1 has an important role to play[27]. In this case, when levels of cholesterol are high enough, both SCAP and HMG-CoA reductase compete in order to bind Insig-1. If HMG-CoA reductase binds to Insig-1, Lys248 (human HMG-CoA reductase) will be ubiquitinated and the protein is then quickly degraded through an ubiquitin-proteasome mechanism[28].

The catalytic activity of HMG-CoA reductase can be modulated by phosphorylation[29]. Next to His 866, one of the active site residues, there is a serine residue (Ser872), which can be phosphorylated. The phosphorylation of this residue leads to the decrease of the catalytic activity of HMG-CoA reductase by decreasing the affinity of the enzyme to NADPH[30]. The position of the serine, so close to the catalytic histidine, is well conserved in superior eukaryotes, which suggests that the phosphoserine can interfere with the ability of histidine to protonate coenzyme A thianion before it is released from the active site[31]. Alternatively, it is supposed that the

phosphoserine can also prevent the closure of a C-terminal region which is thought to trap the substrate in the active site and facilitates catalysis. The subsequent dephosphorylation of the serine completely restores the catalytic activity of HMG-CoA reductase[29].

1.3.2. Statins: the most Common HMG-CoA Reductase Inhibitors

In order to try to diminish the escalading number of deaths caused direct or indirectly by high levels of blood cholesterol, scientists started to investigate the best way to help reduce these levels[32]. They soon found out that controlling the ingestion of cholesterol containing food was not enough to control its concentration in the blood[33] and so they had to resort to other approaches, such as bile-acids sequestrants, nicotinic acid, fibrates and probucol. Still, the efficiency of these medications was limited and the search for better cholesterol lowering drugs ensued. When, in mid-'70s, compactin, the first HMG-CoA reductase inhibitor was discovered by Endo and coworkers, the drug was tested for its efficiency as a cholesterol lowering medicine[34-35]. As the results were favorable, development of new and improved statins rose and soon after it became the most used treatment for high levels of blood cholesterol[36].

Statins (Figure 15) are potent competitive inhibitors of HMG-CoA reductase[37] and can be divided in two types, based on their structure[38]. Type I statins (lovastatin, pravastatin and simvastatin) are natural fungal products and type II statin are fully synthetic. All statins share an HMG-like moiety, covalently linked to a rigid hydrophobic group. When administered, the HMG-like moiety of mevastatin, simvastatin and lovastatin is in an inactive form, which is later hydrolyzed *in vivo* by cellular enzymes (esterases)[39]. Structurally, type II statins are characterized by the presence of larger hydrophobic regions and attached fluoro-phenyl groups[40].

The ability of statins to inhibit HMG-CoA reductase arises from the HMG-like moiety, which competes with HMG-CoA to bind to the HMG bind site of the enzyme. Though the hydrophobic part of statins is different from the coenzyme A portion of the substrate, these non-polar groups also contribute to blocking the access of HMG-CoA to the active site. The affinity of the enzyme for statins is slightly higher than its affinity for the substrate[41].

Even though statins are widely used nowadays as the major drug against high cholesterol levels, there are still some side effects linked to them. These include skeletal

muscle-related toxicity (myalgias, rhabdomyolysis), cataracts, vascular lesions in the central nervous system and testicular degeneration[42-43].

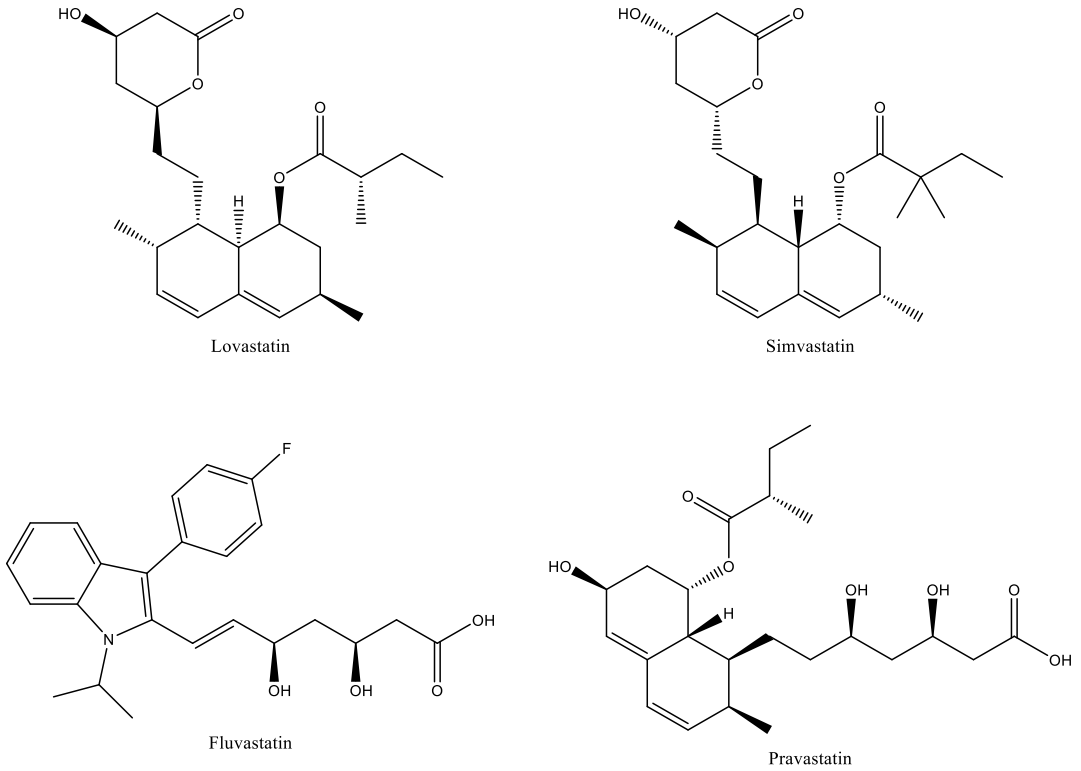


Figure 15 – Structure of some statins.

Since different statins seem to cause different side effects, it has been suggested that these variations are caused by the singular chemical structure of the non HMG-like part. These structural differences can cause them to bind to unwanted targets, alter associated downstream metabolic pathways and produce unsafe side products[44]. While some of these can cause beneficial (pleiotropic) effects (plaque stabilizing, anti-inflammatory and antithrombotic effects, etc.[45]) others can be adverse and even lethal (rhabdomyolysis). Fatalities due to rhabdomyolysis lead to the withdrawal of one of the statins (cerivastatin) from the market in 2001[46]. Still, in spite of the possibly of such serious side effects, this did not preclude atorvastatin from being the most profitable drug in the world over the last 8 consecutive years (with sales over 8 billion Euros each year), which demonstrates the great need of drugs that lower the blood concentration of cholesterol.

CHAPTER 2

COMPUTATIONAL METHODS

2.1. Protein Structure and Model Building

In order for computational biochemists to do their job, they need to have a model of the system they wish to work on. In our case, the objects of our studies were always enzymes. Protein and enzyme models are stored in an online database call Protein Data Bank (PDB)[47]. This is the biggest website were one can go to and search for the tridimensional structure of a given protein, as well as other systems such as viral capsids and mitochondria. With the advance of X-ray crystallography, Nuclear Magnetic Resonance (NMR) and electron microscopy, matched with the progress made in molecular biology, the number of protein structure in PDB has grown exponentially over the last few years (Figure 16).

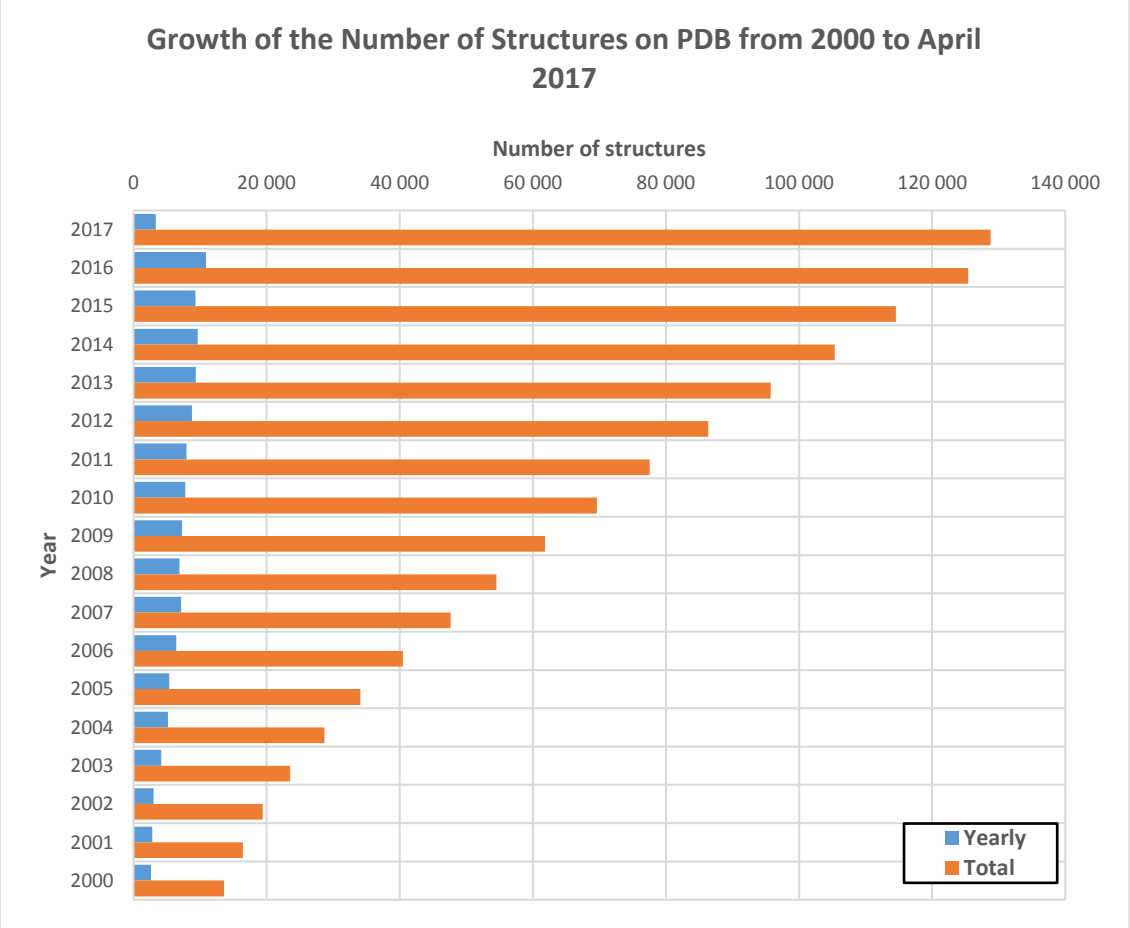


Figure 16 – Graphical representation of the number of structures on the Protein DataBank website, and how it as changed over the years.

Currently there are over 100.000 structures archived in this website (exactly 128.783 at the time this page was written, at the beginning of April 2017) and each week more structures are added. Approximately 89% of these were obtained through the use of X-ray crystallography, 10% through NMR and 1% through electron microscopy, hybrid techniques and others.

PDB structures have a collection of all the atoms of a given object (protein, DNA, RNA, etc.) and their coordinates in the tridimensional space. When opening one of this files, one can view how that molecule is composed and how its atoms are arranged in space.

After choosing which object we are going to study, we should first check PDB (and other similar databases) to see if there are tridimensional structures available for that protein. If multiple structures exist, we must first evaluate each one of them to see which structure best suits our study. Usually we tend to favor structures with better resolution and, if possible, with substrates, products or analogs in the active sites. These compounds give us a lot of information about the configuration of the protein when its active site is full, since it can be very different when there is no molecule in it. Also, the fewer modifications a structure has in relation to the native protein, the better. Finally, one should also be aware that frequently the protein in PDB is not that of the species one wants to study. For the most part we want to study human proteins, but sometimes they are not available. In these cases, we must search for a protein from a similar species (for example, of another mammal) and check if both amino acid sequences are similar, with special focus on the active site residues. If the sequences are very similar and the active site residues are conserved, then we usually assume that both proteins work in the same way, and so the results obtained for one can be extrapolated to the other. In more difficult cases there is always the option of modeling the protein using solved structures with high sequential similarity. This method, however, should be avoided and used only if it is absolutely necessary, because we can never be 100% certain that the final structure resembles, indeed, the actual real structure of that protein. Also, we must conduct further experiments that help us validate the final structures, such as using molecular dynamics to see if denaturation does not occur over time.

Another problem with using protein models, especially when these are obtained through X-ray crystallography, is that some atoms do not appear in the structure. In the case of X-ray crystallography, atoms with low electron density do not show up on the diffraction pattern, which means that their position cannot be assessed. This is more noticeable for hydrogen atoms, which rarely appear in PDB structures. They must be added after the structure is chosen and downloaded from the database.

There are several computer programs and online servers that automatically add missing hydrogens to protein, however such subject is not that trivial. For most atoms, the number of hydrogens and their position can be easily inferred, but there are some cases in which it may be difficult to know in which position an hydrogen must be added (for instance, if it makes a hydrogen bridge or not) and the worst case scenario is when we are not sure whether a residue is protonated or not. For this last case, which is usually common in amino acids with multiple ionization states (histidine, lysine, arginine, glutamate and aspartate), it is important to know the optimum pH of the protein and check (again, using online servers) the pKa for each of these amino acids in the protein environment. Then, depending on the results and, whenever possible, using results already published, we can decide whether or not to protonate a given residue. It is worth noting, though, that these algorithms are not 100% accurate and therefore the results given by them should not be followed blindly but regarded more as suggestions than actual truth about the protonation state of the amino acids. Oftentimes visual inspection is also a good tool to help ascertain which protonation state is more likely.

Sometimes mutations are added to the protein prior to crystallization, and these are the residues that ultimately appear on the structure. This is done for various reasons, sometimes to help the protein crystallize or to prevent catalysis from happening, which means that the substrate becomes trapped on the active site and we can examine the way it binds to the protein. Mutated residues must be modified into the wild type ones too before using the model. For small modifications like this, it is easy to revert the amino acid back to the one encountered in the wild type enzyme by simply using an editing software, such as GaussView[48], and modify the atoms one by one.

Another common problem that exists in downloaded structures is when there are residues of the protein missing. This is due to the way X-ray crystallography works: some unordered parts of the protein (most frequently, loops) might not crystalize correctly and, therefore, do not have a diffraction pattern that allows for their unambiguous resolution. These missing parts must be added subsequently to the model. This case is a bit harder to address than the one before, since the parts missing can be long. In this case, a better approach is to superimpose the structure with another that has the missing parts in it and simply copy and paste the atoms to our model. If no other similar protein is available, we can try modeling it using protein with a similar structure. In either of these cases, it is wise to follow up with a small molecular dynamics to try and see if the protein does not lose its integrity.

2.2. Molecular Mechanics

After we have completed the model of the object we want to study, be it a protein, DNA, protein-ligand complex, or any other option, then we can finally start researching several different things about it. Computational biochemistry is such a vast area, with so many different techniques, that nowadays one can study virtually anything, especially with the fast development of computers and processors. One of the most versatile computational methodologies that enables us to analyze and calculate different characteristics of a protein is molecular mechanics[49-50].

Molecular mechanics is a means to simplify a system of atoms and calculate its different properties using classical (or Newtonian) mechanics. It describes the model using a force field, which is a combination of a general equation for the energy as a function of the atomic coordinates, and a set of parameters that are needed to adapt the equation to the specific molecular system under study. The development of molecular mechanics began when computers were not powerful enough to treat a very large system with quantum mechanics. These calculation required very high computational power, which was not available at the time. With molecular mechanics, these systems can be described using the laws of classical mechanics. In molecular mechanics the smallest unit we can treat is the atom, however if we want to speed up our calculations we can form units composed of united atoms or even whole residues. These approaches allow us to speed up our calculations at the cost of atomic precision. With the ever growing computational power, today we do not need to perform these types of calculations for the systems we normally study. Still, in cases where we have a protein with a very high molecular weight, the use of these coarse graining methods can be helpful. In this work, we will focus only on atomistic models, since our models were manageable in a reasonable time frame using this approach.

2.2.1. Force fields

As it was stated previously, force fields are an essential part of molecular mechanics. They are used to derivate the total energy of the system in a given conformation. Several types of force fields exist and are used depending on the type of system and simulation time. In this work, we used exclusively the AMBER (Assisted Model Building with Energy Refinement) force field[51] to characterize proteins. For ligands we used GAFF[52] (General AMBER Force Field) to describe them, since this force field

has all parameters to characterize almost all organic molecules composed of C, N, O, H, S, P, F, Cl, Br and I.

A force field is basically a mathematical expression and a set of parameters that enable the description of the dependence of the energy of a system on the coordinates of its particles. Most molecular systems are too complex to be treated with quantum mechanics, so we resort to using simple classical functions that represent it as a group of atoms connected by harmonic (elastic) forces and a set of parameters obtain through either experimental (X-ray and electron diffraction, NMR, spectroscopy, etc.) data or quantum calculations on smaller models. Although there are currently many different force fields in literature, which are used depending on complexity of the system and the level of accuracy we want to obtain, typical they are defined using the following expression:

$$V_t = \sum_{b=1}^{n-bonds} V_{bonds} + \sum_{a=1}^{n-angles} V_{angles} + \sum_{d=1}^{n-dihedrals} V_{dihedrals} + \sum_{p=1}^{n-nbpairs} V_{ee} + \sum_{p=1}^{n-nbpairs} V_{LJ} \quad (1)$$

In which the first term is the sum of the potential energy associated with bond stretching, the second corresponds to angle bending, the third to dihedral torsions and the last two relate to the non-covalent interactions, electrostatic and Van der Waals interactions, respectively.

In molecular mechanics, bond stretching is often represent as a harmonic function. Covalent bonds are described as springs that either elongate or shorten and therefore their potential energy can be calculated using Hooke's Law. This means that, in the force field general equation (equation (1) the term V_{bonds} takes the following form:

$$V_{bonds} = k_r (r - r_{eq})^2 \quad (2)$$

Where k_r is the bond force constant, r_{eq} is the length of the bond in equilibrium and r is the current bond distance. The force constant is inferred from infrared or Raman spectra, while r is obtained from either X-ray diffraction experiments or quantum mechanics calculations.

One of the major drawbacks of molecular mechanics is the fact that it cannot deal with breaking and formation of new bonds, and we are therefore unable to study chemical

processes. This is due to the fact that this approximation is incorrect for bond displacements bigger than 10%, which is a problem that derives from the fact that the in a harmonic potential the energy keeps growing as the bond length increases. If we were to use, for example, the Morse potential (or some other which, as the distance increases, the energy levels), the breaking of chemical bonds could be simulated. However, the parameterization for these types of potential is rather complex and must be done for each individual case, while the results are not always accurate. For this reason, most force fields use a harmonic function instead, since for near equilibrium conformations, the results obtained with these potential are acceptable.

The energy associated with angle distortion is also estimated using a harmonic approximation. Therefore, we can calculate V_{angles} using the following expression:

$$V_{angles} = k_\theta \, (\theta - \theta_0)^2 \tag{3}$$

In this case, k_θ is the force constant, θ_0 is the angle at the equilibrium configuration and θ is the current angle. The quadratic approximation is fairly good to calculate the potential energy of angles, but it has also two weaknesses: it is not accurate for calculating energies when the angle is very different from the standard configuration and it does not account for cases in which a molecule can have several equilibrium angles, as it happens it many organic molecules. Nevertheless, this approximation is used rather frequently and it presents good results in most of the cases.

For dihedral torsion, it becomes more difficult to establish an expression that can accurately calculate the potential energy associated with this type of interaction. Torsional energy relates to the energy variation associated with the rotation of the bond between two atoms, B and C, in the four atom sequence ABCD, which are sequentially bound. The problem with dihedrals is that they are periodic, and so, in order to calculate their energy we must use the ensuing expression:

$$V_{dihedrals} = \frac{V_n}{2} \, [1 + cos(n_n \, \phi - \gamma)] \tag{4}$$

V_n represents the height of the energy barrier, n is the number of minima or maxima that exist between 0 and 2π, ϕ is the current dihedral angle, and γ is the phase. Torsion parameters are often calculated using *ab initio* methods and then refined with the help of experimental approaches, such as vibrational spectra. In macromolecules, such as proteins, where there is a large number of dihedrals, these motions must be defined more precisely than angle bending or bond stretching, since dihedrals are necessary to ensure the correct degree of rigidity of the molecule and to reproduce the major

conformational changes. Some force fields also employ an extra variable to account for improper torsions, which corresponds to the energy associated with the rotation of a bond in which the four atoms analyzed are no necessarily bond sequentially. Improper torsions are important to ensure the planarity of sp^2 atoms.

As for the non-covalent interactions, we have two different terms, one for calculating the potential energy associated with electrostatic interactions, and another for Van der Waals interactions.

Electrostatic potential energy is related to the electronegativity of the different atoms that compose a molecule. Electronegativity is a chemical property that represents how much tendency an atom has to pull an electron towards itself. Therefore, atoms with more electronegativity tend to pull electrons from other atoms of the molecule, thus creating an uneven distribution of electrons in chemical bonds and the formation of an electric dipole. The electrostatic potential energy is calculated by assigning a specific partial charge to each nucleus and using Coulomb's law, which states that the electrical force between two particles depends on the magnitudes of the charges of each particle and on the distance between them. The relationship between all this quantities can be expressed by the following mathematical expression:

$$V_{ee} = \frac{q_1 q_2}{4\pi\varepsilon_0 r} \tag{5}$$

in which q_1 and q_2 represent the partial charges of each atom, ε_0 is the vacuum permittivity and r the distance between both atoms. Through this expression we can see that the energy associated with the electrostatic potential is directly proportional to the partial charges of the atom and inversely proportional to the distance between then, meaning that the further apart the atoms are the smaller (in absolute terms) the energy associated with this potential.

The last contribution that is usually taken into account when calculating the energy of a system is that of the Van der Waals interactions. In order to accurately simulate the behavior of pairs of atoms at a certain distance, the equation used to calculate the Van der Waals energy must account for two things: the repulsion felt by the atoms when their respective orbitals superimpose (due to Pauli's exclusion principle) and the attraction created by London dispersion forces, in which the instantaneous uneven distribution of electrons in an atom can create a temporary dipole which, in turn, create another induced dipole in a nearby atom. Van der Waals forces vary very quickly with the distance between both atoms being very repulsive when they are really close. However with the increase in the distance between them, the atom start to attract each other, until the point

when they no longer feel one another. This behavior can be described mathematically through the Lennard-Jones potential (also known as the 6-12 potential):

$$V_{LJ} = 4\varepsilon\left[\left(\frac{r_m}{r}\right)^{12} - \left(\frac{r_m}{r}\right)^{6}\right] \quad (6)$$

In this expression ε represents the minimum potential energy reachable, r_m is the distance at which this minimum potential is reached and r is the current distance between the atoms. The Lennard-Jones potential is rather good at simulating the Van der Waals forces felt between two atoms since the positive factor (power 12) models the repulsion, the negative factor (power 6) models the attraction and the first can be obtain by just squaring the second, which saves computing time.

Most force fields use only these five terms in order to calculate the total energy of the system. However, depending on the force field, other terms might also be employed in addition to the one described. These are called cross-terms and usually vary depending on the force field, and can describe, among other properties, the coupling between bending, stretching and torsion energies. These terms are often added to correct the calculated energies and render results more accurate.

For this work we used mainly the AMBER force field. AMBER is an “all-atom” force field, which means that the smallest unit it considers is the atom. With this force field, all of the atoms of the system are explicit and treated according to their place on the structure. However, protons and electrons are implicitly taken into account by parameters that describe the atoms (atomic charge and van der Waals radius).

2.2.2. Energy Minimization

When performing energy minimization calculations, one tries to reach the structure of the system that has the smallest possible energy, i.e. the one which correlates to a minimum on the potential energy surface[49-50]. In proteins it is rather difficult to find the absolute minimum, since it would require a very large amount of time to explore the whole of the potential energy surface, which in the case of proteins, due to the many degrees of freedom that it contains, is relatively big. For this reason, when minimizing finding a protein structure with minimum energy we have to settle for a local minimum. Furthermore, for proteins there are a lot of degenerate local minima, which means that at room temperature, a protein is never found at an absolute minimum, quite the opposite, since in the case of proteins we can have different local minimum with an

abundant population. Therefore, using a local minima as a starting point for the following calculations is sufficient.

AMBER employs two different algorithms to search for a local minimum structure: steepest descent and conjugate gradient. The steepest descent algorithm is useful when the geometry of the system is very far from a local minimum and it is ensured that it will converge to a minimum. As the coordinates of the system are changed, the gradient (the derivative of the potential energy) is calculated, and the algorithm moves the system in the direction that contradicts the gradient. It the case of a one dimensional system, the algorithm takes the following form:

$$x_{i+1} = x_i - \varepsilon \nabla V(x_i) \tag{7}$$

where x_{i+1} is the new position of the system, x_i is the previous position, ε is the magnitude of the step taken towards the minimum and $\nabla V(x_i)$ represents the magnitude of the potential energy function at the previous position. One of the flaws of this method is that, since it uses only the current gradient, the direction followed by the path tends to be somewhat erratic, changing with each iteration and creating a zig zag pattern. This pattern is particularly noticeable when the system is already close to a minimum. Since this weakness translates into a loss of performance during the minimization, AMBER also employs the conjugate gradient algorithm. The difference between both algorithms is that, contrarily to steepest descent, conjugate gradient uses not only the gradient from the current step but it conjugates it with gradients from other previous steps. For this reason, the path followed by this algorithm is much more focused after some iterations, which allows for convergence to be reached in very few steps after a certain point.

2.2.3. Molecular Dynamics:

The minimization of a system allows us to detect a structure in which it will be more stable (have more negative potential energy). However, most of the times, we do not wish to find out which conformation is the most stable, but rather see how the system evolves as time goes by and which set of enzyme conformations are the most occupied at physiological temperature. In order to achieve this, we must employ molecular dynamics, in which the potential energy function is used together with the Newton equations of motion[49-50, 53]. The conjugation of both allows us to simulate the behavior of our molecular system through the course of a time lapse.

To “kick-start” our model into motion, we must first assign random initial velocities to each atom, which is done according to a Maxwell-Boltzmann distribution, and taking into account the expected temperature of the system. At this stage, we need to calculate the acceleration for each particle in order to know their position after a certain time has elapsed. From Newton’s second law of motion we know that:

$$F_i = m_i a_i \tag{8}$$

which means that the force acting upon an atom *i* of the system is equal to the product of the mass of that atom (m_i) and its acceleration (a_i). The acceleration can also be represented as the second order derivative of position with respect to time, meaning that another way to represent a_i is:

$$a_i = \frac{d^2 r_i}{dt^2} \tag{9}$$

Combining equations (8 and (9 we get a final expression, which is used in molecular dynamics to calculate the trajectory of each particle of the system.

$$F_i = m_i \frac{d^2 r_i}{dt^2} \tag{10}$$

Each atom will move according the force *F* being applied to it. In order for the trajectory to be calculated along a time scale, we need the initial positions of the atoms, which can acquired by X-ray crystallography, NMR and other methods. The most difficult term to calculate in molecular dynamics is the acceleration, which cannot be calculated analytically for systems bigger than two atoms, since it is dependent on all atoms of the system. The solution found for this problem is to solve it numerically, which can be done using one of several integration algorithms. AMBER uses the Leap-frog algorithm, which takes the following form:

$$v\left(t + \frac{1}{2}\delta t\right) = v\left(t - \frac{1}{2}\delta t\right) + a(t)\delta t \tag{11}$$

$$r(t + \delta t) = r(t) + v\left(t + \frac{1}{2}\delta t\right)\delta t \tag{12}$$

With the Leap-frog algorithm, the velocities of the particles are calculated based on their velocities at half-step behind ($v(t + ½\,\delta t)$) and their acceleration at time *t*. After

these velocities are calculated, their positions can be obtained using equation (12. The name for this algorithm comes from the fact that the positions and velocities are calculated at interleaved time points in such a way that they “leapfrog” over one another.

2.2.3.1. Molecular Dynamics Parameters

In molecular dynamics one of the most critical parameters to choose is the time step. At each time step, the new positions and velocities for each atom are determined from the old ones and the forces being applied to them, so in order to take the least computational time possible it would make sense to use a large time step which would allow us to scan the movements of the system in a large temporal range while doing the least number of calculations possible. The problem with this approach is that calculating structures which are temporally far apart can lead to wrong results, since the new atom positions and velocities are calculated from the ones of the previous step. If the forces are relatively constant during the step, then the results should be accurate, however with a large time step we cannot be certain if this is the case, and the calculations will result in wrong potential energies and positions for the atoms. To avoid these kinds of errors, it is customary to select a time step that is ten times smaller than the fastest vibration in the system. In most simulations, these vibrations are the stretching vibrations of the hydrogen atoms, which have periods of about 10 fs, meaning that the time step should be 1 fs. In order to increase the time step, the SHAKE algorithm[54] implemented in AMBER can be used, which freezes these vibrations. The second quickest vibrations are the stretching of C-C bonds. These have a period of circa 20 fs, which means that we can increase the time step to 2 fs.

It is common practice to add explicit water molecules to the model when studying a system through molecular dynamics. TIP3P[55] is a water model implemented in the CHARMM force field is one of the most frequently used. It is characterized by three point charges, one for each of the atoms that compose the molecule, but only one van der Waals radius, corresponding to the oxygen atom. Another characteristic of the TIP3P water type is that it does not have an angle parameter; instead there is a bond parameter between the two hydrogen atoms. These simplifications made to the water molecules may seem irrelevant, but they play a crucial role for the simulation. When explicit solvent is used, the number of water molecules can surpass that of protein atoms by more than ten-fold, which would mean that most of the simulation time would be spent calculating the behavior of these molecules. In the case of TIP3P waters, the number of interactions calculated during the simulation is decreased significantly, simply by eliminating these

two van der Waals parameters. Also, by substituting the angle parameter by a bond parameters, the calculation time is reduced, since calculating angles is computationally more demanding than calculating bonds.

Besides the added calculation time needed when using explicit solvent, there are other complications that arise from this approach. For instance, the fact that the structure of the solvent is not as cohesive as that of the protein itself. If we were to run a normal simulation with the water molecules simply surrounding the protein, it would result in the loss of some of those molecules, and the distortion of the water cap. Another problem comes from the fact that those water molecules closest to the solvent-vacuum interface would not be properly feeling the rest of the solvent, which would lead to some incorrect results. For this reasons, when using explicit solvent we also employ periodic boundary conditions, an elegant solution to deal with the problems described above. With periodic boundary conditions, the simulation cell is simply repeated an infinite number of times in all three coordinates of the system. The most common shape of the box is a cube, however other shapes can be used if one wishes, for example, to reduce the number of solvent molecules in the system.

During the molecular dynamics, it is possible to choose one type of ensemble conditions from several that are available to the user. The most commonly used amongst them are the microcanonical ensemble (NVE), the canonical ensemble (NVT) and the isothermal isobaric ensembles. When using the NVE ensemble, the number of particles in the system, the volume of the solvent box and the energy are all kept constant throughout the simulation. On the other hand, using NVT conditions allows the energy of the system to change while the number of particles, volume of the box and temperature of the system are kept constant. Finally, in NPT conditions it is the temperature, pressure and number of particles that are kept constant. From the three ensembles presented, the NPT is the most similar to a physical aqueous solution in room conditions, and for this same reason it is the one most frequently used in molecular dynamics. However, the canonical ensemble is still useful in the first stages of the simulation, when we want to optimize the molecules of the solvent, since after they are added to the system they are not in a relaxed and optimal geometry. If the simulation is run at this point with in NPT conditions, the system will most likely expand and form cavities in the solvent, due to the fact that the pressure is changeable. Since volume is kept constant in the canonical ensemble, the water molecules are able to acquire a relaxed conformation without deforming the solvent box.

During the molecular dynamics simulation, the temperature is kept constant through the implementation of a thermostat. These are simply functions which, in some

way, alter the velocities of the atoms that are part of the system. Changing these velocities simulates an increase or decrease in the temperature of the system, which can then be adjusted to the designated value. The thermostat used in Amber is the Langevin dynamics algorithm, in which stochastic collision of an adequate force are simulated upon the atom. What is expected with this approach is to model the effects of the surround environment in the system implicitly.

There are, however, some issues that arise from the usage of periodic boundary conditions. One of the most problematic is the fact that, since theoretically the system is infinite, there is also an infinite number of contributions that must be calculated in order to obtain the potential energy of the system. Some terms are passive to be calculated in only one cell (such as those relating to bonds, angles and dihedrals, which are the same in all boxes). There are, nonetheless, terms relating to pairwise non-bonded interaction that occur between molecules present in one cell and all the other molecules in the repeated cells, up to infinity. In order to avoid these infinite calculations, we must employ some approximations.

We can effectively reduce the number of van der Waals interactions that must be calculated in one of the easiest ways possible: simply by limiting the number of atoms that interact with each other. This can be done by introducing a cut-off radius around each atom, this way the central atom will only interact explicitly with those that are found within this distance. This approximation can be easily understood if we take into account what was stated earlier: that the energy associated to the van der Waals interactions decreases very rapidly as distance increases. However, it is also true that with the distance the number of interaction also increases, which means that at large distances it is still expected that the contribution of these interactions is still meaningful. Since van der Waals interactions are always addictive to the total energy, we can simply add a correction factor that accounts for the density of the system, which also considers that it is homogeneous after the cut-off distance.

A different approach must be used when it comes to calculating the contribution of electrostatic interactions. For these we employ the Particle Mesh Ewald method, in which they are divided into large and small range interactions. Long range interactions, which are considered for atoms at a distance larger than 10 Å (this distance is usually the same as the cut-off used in the calculation of the van de Waals interactions) are taken into account through a summation in Fourier space, whereas small range interactions are calculated using the force field equation. The reason why we need a different method to calculate the contributions of electrostatic interactions is that, contrarily to the van der Waals interactions, these decay in an inverse fashion, quadratic to the force.

When building the solvent box around our protein some simple rules must be taken into account, especially its size. It is important that the cell is large enough that it hinders interactions between two proteins. We want to keep the proteins from seeing one another because the simulations are done with the purpose of studying them at low concentrations in the solution. For this reason, the size chosen for the solvent box is usually that of the cut-off distance for non-bonded interactions, since past this distance the interaction are no longer considered explicitly.

2.2.3.2. Parametrization of Ligands and Non-standard Residues

As stated previously, in order to run a molecular dynamics simulation, we first must have a parameter file for all atoms of the system. This parameters are dependent on several factors, including atom type and which other atoms it is bound to, amongst others. If we search through the literature, it is possible to find parameters for a large variety of biomolecules; however this list is far from complete. For the most common of biological system, such as proteins, nucleic acids, phospholipids and carbohydrates, these parameters do exist and are already well documented. But for more specific molecules, like substrates or inhibitors, these are most probably not yet available on the literature, and so it comes down to the person in need of them to calculated these new set of parameters from high level quantum mechanics methods.

For bond and angle parameters, they are determined by scanning the proper coordinates of the molecule. An energy optimization is initially performed in order to calculate the equilibrium length and force constant of a bond. Subsequently, this same bond is stretched and shrunk, so that an energy profile that should be roughly quadratic near the minimum is obtained. The two parameters that most authentically mimic this energetic profile are chosen. This protocol is usually only put in practice for out of the ordinary systems, such as metallic centers. For molecules that are most common in biological systems, it is actually much easier and faster to just attribute new parameters for bonds and angles based on structure similarities with other molecules that are already parameterized. We first define the atom type of each atom of the system, and then parameters are attributed in accordance with GAFF. The parameters for dihedrals are oftentimes ignored, since they are difficult to calculate and rather numerous. Furthermore, 1-4 interactions are already taken into account approximately by non-bonded interactions.

As for the atomic charges, these, however, cannot be assigned by similarity and are dependent on each atom type and the environment in which the molecule is inserted, which also changes its polarization. The protocol for calculating the charge parameters begins with the minimization of the molecule while it is inserted in the site of interest. Usually a QM/MM calculation with electrostatic embedding is preformed, where the high layer corresponds to the molecule to be parameterized and the low layer to the enzyme. After the minimization is finished, we employ the RESP (Restrained ElectroStatic Potential fit) method[56] in order to assign the charges to each atom. In a RESP calculation the charges are assigned to mimic the electrostatic field cleated by the molecule in the QM model. The method is called “restrained” because one can make use of certain restrictions when attributing the charges. This can be helpful in some cases, for example, when there are equivalent atoms in the molecule that have the same charge or for imposing a certain charge on an atom.

It is also important to bear in mind the conformation of the molecule when performing ESP calculations. If the molecule is minimized in vacuum, its configuration might be quite different from the one it takes when inserted into the enzyme. Some groups might interact with one another and that might interfere with the calculation of the charges. For example, if a molecule as opposing charges at either end and is somewhat flexible, this groups can interact with each other, which would cause the charge calculations to be wrong. It is for this reason that the ESP calculation should ideally be done using an ONIOM model, with a high layer that consists of the ligand and a low layer that refers to the protein. This way, the charges calculated for the ligand will be perfectly in tune with the conformation it adopts when inserted into the enzyme. The limitation of the approach is that in case we wish to study a different configuration of the same molecule (for example, if it is inserted on a different environment, or if it is in solution) the charges calculated will not be correct, and a new ESP calculation might be needed.

2.2.3.3. Relaxation of Structures

Proteins are very complex structures that can contain a large amount of atoms, each of which passible of having an important function for the overall structure. For this very reason, when studying enzymes and enzymatic catalysis, it is rather important to make sure the structure of our model is in a state similar to that it is expected to have in its natural conditions. Ensuring the protein is in such a state is one of the most common applications of molecular dynamics in this field. With this method, we expect to stabilize the structure of the enzyme, or enzyme-substrate complex, after we perform

modifications on the original structure. The need to modify the structure that comes from crystallography comes from the fact that it seldom is in the state we intend to study the enzyme. More often than not, the structures do not have the right substrate (or they have no substrate at all) in the active site, or some mutations has been made on the protein. Oftentimes the proteins are not fully crystalized and we need to model parts of it as well. In order to carry on with the study of the mechanism, we need to arrange the protein so that its structure matches what we are trying to study. Often only minor modifications are necessary, such as the transformation of one substrate or amino acid into another, but sometimes bigger transformations are needed, especially when parts of the protein are missing. In the case of only small modifications, a simple structure minimization might be enough to ensure the system is in the correct configuration. However, in the case of large modifications, in order to stabilize the structure in the right configuration, larger movements of the atoms are required and a simple minimization might not suffice, and so it is usually accompanied by a molecular dynamics simulation.

When running a simulations, the protocol used is usually rather simple and straightforward. First, when preparing the input file, the structure is checked to see if there are any mistakes. Then, solvent molecules (generally water) are add so that they form a box around the protein, leaving at least a 12 Å margin from the surface of the protein to the edges of the box. Next, we add counter ions, either sodium or chloride, to the system, so that the total charge equals zero, which is a standard requirement of the Particle Mesh Ewald method. One should always ensure that, when adding counter ions, these do not fall near the active site, otherwise they can disrupt the course of the dynamics and the structure of the protein. After all these modifications, the initial structure is ready, however before commencing the dynamic simulation, it is usual to run an energy minimization calculation using the following protocol: first, the hydrogen atoms are minimized, after which follows any substrate or residues that might have been modeled. Next, the solvent molecules are minimized and lastly all the whole system. In the end of all these minimization calculations, the molecular dynamics can start, first with the heating of the system at constant volume (canonical ensemble), and then finally the production phase takes place with an isothermal isobaric ensemble.

An important question one should ask when doing any dynamics is how long should it be ran for. This is a rather delicate matter that must be met with care. The structure obtained through crystallography is no more than an average of the structures found in the crystal. That being said, it is understandable that those structures found most frequently will have a bigger contribution to the final structure of the protein, whereas those which are found with less frequency having a smaller contribution. What

we obtain in the end is a Boltzmann distribution of the system, which is encased in a single structure. When performing a molecular dynamics simulation, this correlation between the crystallographic structure and the Boltzmann distribution is destroyed, and so it is usual to keep the calculation running for the minimum amount of time possible, only that required to relax the modified sections of the model. We would hope that keeping a small simulation time would not disturb the properties of the system. However, as many studies have shown, even a very small simulation can disrupt the initial configuration of the system, and this might affect both the pathways and energies that are resultant from mechanistic studies.

After the dynamics, in order to assess whether the modeled part of the system is already relaxed or not, we perform an analysis of the root mean square deviation (RMSd). If the model is relaxed, we can then use the last structure to build the model that will be used in the subsequent studies, or if we find it yet needs to undergo molecular dynamics for a bit longer we continue the calculation from this last point.

2.3. Quantum Mechanics

Quantum mechanics began its history at the start of the 20th century, when the earliest molecular and atomic systems were first studied[57-60 a]. In the beginning, physics tried to apply the laws of Classical Mechanics to said systems, only to find that they did not always followed these laws, or that they presented certain limitations. Thus, quantum mechanics emerged from the need to come up with new theories that could explain the results that were obtained when studying atomic scale systems.

At the core of quantum mechanics sits the Schrödinger equation:

$$\hat{\mathrm{H}}\Psi = E\Psi \tag{13}$$

By solving this equation, we can, theoretically, solve all properties of a system, with special regard to its energy. These properties are derived from the motions of electrons and nuclei of each atom. For ground-state chemistry studies, such as the enzymatic reaction discussed later, the time-independent Schrödinger equation is sufficient (equation (13). What this equation states is that if we apply the Hamiltonian operator ($\hat{H}$) to a certain wave function (Ψ) and the result obtained is itself proportional to Ψ, then we

[a] QM theory is well presented and organized in any of these books, hence the lack of specific references in the following sections.

can conclude that Ψ is a stationary state and an eigenstate of H and the proportionality constant, E, is the energy of that same state.

One of the problems of using the Schrödinger equation is that it is analytically unsolvable for systems with two or more electrons. For this same reason, in order to calculate energies for these types of system several approximations must be made.

2.3.1. Wave Function Methods

All properties of a system can hypothetically be derived from the motions of the electron and nuclei that compose it. Using the Schrödinger equation it is theoretically possible to describe these motions and from them calculate the properties of the system. One of the terms of the Schrödinger equation is the Hamiltonian operator ($\hat{H}$).

The Hamiltonian operator includes different terms: those associated with kinetic energy, i.e. the movement of the particles, and those associated with potential energy, i.e. the Coloumb interactions between the various particles. For a general N-particle system, $\hat{H}$ becomes the following equation:

$$\hat{H} = \hat{T}_n + \hat{T}_e + \hat{V}_{ne} + \hat{V}_{ee} + \hat{V}_{nn} \tag{14}$$

Where $\hat{T}_n$ is the kinetic energy of nuclei, $\hat{T}_e$ the kinetic energy of electrons, $\hat{V}_{ne}$ the attraction felt between nuclei and electrons, and finally $\hat{V}_{nn}$ and $\hat{V}_{ee}$ represent the repulsion felt between nuclei and electrons respectively. Each of these terms, in turn, can also be represented in such a way that the $\hat{H}$ takes the following form:

$$\hat{H} = -\sum_K \frac{1}{2M_K}\nabla_K^2 - \sum_i \frac{1}{2}\nabla_i^2 + \sum_{K>L} \frac{Z_K Z_L}{\left|\bar{R}_i - \bar{R}_j\right|} - \sum_{i,K} \frac{Z_K}{\left|\bar{r}_i - \bar{R}_K\right|} + \sum_{K>L} \frac{1}{\left|\bar{r}_i - \bar{r}_j\right|} \tag{15}$$

$$\nabla_i^2 = \left(\frac{\partial^2}{\partial x_i^2} + \frac{\partial^2}{\partial y_i^2} + \frac{\partial^2}{\partial z_i^2} + \right) \tag{16}$$

For any system with two or more electrons, this equation becomes just too complicated to be solved exactly using the knowledge available to us. In order to reach an approximate result, some simplification have to be introduced.

The Born-Oppenheimer approximation is based on the fact that electron are much smaller than the nuclei, and that they also move at much faster velocities. We can therefore assume that when the nuclei assume a certain position, the electrons will adopt

the arrangement with the lowest energy before the nuclei move significantly. In practice, what this means is that we can separate the Hamiltonian in two different contributions, one part for describing the electronic wave function and another for the nuclear wave function. Solving the electronic Schrödinger for different nuclear arrangements leads to a potential energy surface (PES), the minima of which determines the equilibrium geometries of a molecule.

$$\hat{H}_{total} = \hat{H}_e + \hat{H}_n \quad (17)$$

Electronic Contribution

$$\hat{H}_e = \hat{T}_e + \hat{V}_{ne} + \hat{V}_{ee} \quad (18)$$

Nuclear Contribution

$$\hat{H}_n = \hat{V}_{nn} + \hat{T}_n \quad (19)$$

Even with the Born-Oppenheimer approximation, the electronic Schrödinger equation remains too complicated to be solved for systems with more than one electron, due to mathematical limitations. The main problem with its resolution was the electron-electron repulsion term. Further approximations were thus needed to solve it. In order to deal with this problem, the Hartree-Fock (HF) method, also called the self-consistent field (SCF) method, was developed, which is based on the independent particle model, where each electron is considered to interact with a constant electronic field created by all the other electrons. Each electron is thus related with a one-electron wave function (called molecular orbital, MO) which is the combination of a spatial function that depends on the coordinate of the electron, and a spin function that depends on its spin. The wave function has to satisfy the antisymmetry principle, and must change sign if the coordinates of the two electron are interchanged (Pauli principle). An easy way to build a wave function that respects this principle is by using a Slater determinant of N one-electron orbitals (N is the number of electrons). With this, the N-particle problem is transformed to a set of one-particle problems:

$$\hat{f}_i \chi_i = \varepsilon_i \chi_i \quad (20)$$

where $\hat{f}_i$ is an effective one-electron operator (Fock operator), in which the electron-electron repulsion is treated as an average. χ_i is the corresponding eigenfunction (i.e. MO) and the electron in the MO has the orbital energy ε_i.

The HF equation is non-linear and must be solved iteratively. This is due to the fact that the Fock operators are not independent from one another, and so, in order to calculate the energy it is necessary to know the wave function of all electrons. We employ, therefore, an iterative method in which we begin by selecting a set of molecular orbitals that seem appropriate to the system. The Fock operators for each electron are formulated using this initial estimate, and the HF equation can be solved, from which results a new set of orbitals. This new set is now going to be used to formulate different set of Fock operators, and the whole process repeated until the energy difference between two consecutive iterations is smaller than a set value.

The independent-particle model results in an inherent limitation of the HF method, since the motion of all electrons are correlated in a real system. Neglecting correlation energy (the difference between the HF energy and the exact non-relativistic ground-state energy within the Born-Oppenheimer approximation) leads to large deviations from experimental results, which makes the HF method very poor for exploring chemical reactions. A number of approaches to correct this weakness, collectively called post-Hartree-Fock methods, have been developed to include dynamic electron correlation. For example, Møller-Plesset perturbation theory[61] treats correlation as a perturbation of the Fock operator. Configuration interaction (CI)[62] and couple cluster (CC)[63] are also methods that allow to recover most (in the limit, all) of the dynamic correlation energy. These approaches improve the level of accuracy but become computationally much more demanding, and thus are only suitable for relatively small systems. To investigate large systems, such as enzymatic reactions, a less costly method is needed.

2.3.2. Density Functional Theory

In order to deal with the limitations of the HF method and the computational cost of the post-HF methods, a new alternative was devised and the density functional theory (DFT) approach was born. DFT methods are an alternative to the wave function methods described previously. The basis of DFT is the Hohenberg-Kohn theorem, which shows that the total energy of a non-degenerate ground state is a unique functional of the electron density of the system, namely $E = E[\rho(r)]$. This implies that all properties of the system can be deduced from the ground-state density and the determination of the complicated many-electron wave function can thus be avoided. However, a fundamental difficulty emerges here, that is, the exact functional, i.e. the dependency of the energy on the given electron density, is not known. Various approximations and attempts have been made.

The energy functional can be expressed as follows:

$$E[\rho] = T[\rho] + E_{ee}[\rho] + E_{en}[\rho] \tag{21}$$

where *T* is the kinetic energy, E_{ee} the electron-electron repulsion, and E_{ne} the nuclei-electron attraction. In 1965, Kohn and Sham contributed a significant development, i.e. the orbital-based scheme, in which independent particles move in an effective potential (the non-interacting one electron orbitals are called Kohn-Sham orbitals, ϕ_i). The real system of interacting electrons can thus be described through a system of non-interacting particles by expressing the electron density as a sum of the squared orbitals. Therefore, the total kinetic energy (*T*) is divided into two parts, the kinetic energy of an N electrons non-interacting system (*Ts*) and a missing fraction (*Tc*) relative to the real interacting system:

$$T[\rho] = T_s[\rho] + T_c[\rho] \tag{22}$$

The functional of electron-electron repulsion ($E_{ee}[\rho]$) can be divided into the classical Coulomb interaction (J) and a non-classical part containing correlation and exchange ($E_{ncl}[\rho]$):

$$E_{ee}[\rho] = J[\rho] + E_{ncl}[\rho] \tag{23}$$

The total energy can be written as:

$$E[\rho] = T_s[\rho] + T_c[\rho] + J[\rho] + E_{ncl}[\rho] + E_{ne}[\rho] \tag{24}$$

Then, a definition is done by combining the missing part of the kinetic energy ($T_c[\rho]$) and the correlation and exchange part ($E_{ncl}[\rho]$) to form the exchange-correlation functional ($E_{xc}[\rho]$). The total energy can finally be presented as:

$$E[\rho] = T_s[\rho] + J[\rho] + E_{ncl}[\rho] + E_{xc}[\rho] \tag{25}$$

The first three terms can be calculated explicitly. All the problems have now been centralized in how to accurately describe the exchange-correlation term, $E_{xc}[\rho]$, which encompasses all the unknown contributions to the total energy.

Since the true density is that which corresponds to the lowest energy, the variational principle can be used to calculate the energy. Using it together with the normalization constraints, minimizing the total energy of a determinant constructed by Kohn-Sham orbitals results in the Kohn-Sham equations (similar to the Hartree-Fock equation):

$$\hat{h}_{ks}\phi_i(r) = \varepsilon_i\phi_i(r) \tag{26}$$

where $\hat{h}_{ks}$ is the one-electron operator and depends on the electron density. If the exact form of the $E_{xc}[\rho]$ functional in known, the exact total of the many-electron system can be obtained by iteratively solving the equation (26. The accuracy of a DFT method lies in how accurate the form of $E_{xc}[\rho]$ is.

Many exchange and correlation functionals have been and are still currently being developed. A significant improvement to the accuracy of DFT came from the introduction of the gradient and higher derivatives of the electron density in the functional, i.e. $E_{xc}[\rho, \nabla\rho]$. Later, Becke's introduction of the exact Hartree-Fock exchange as a part of $E_{xc}[\rho, \nabla\rho]$ successfully leads to the popularity of DFT. The DFT methods including HF exchange are referred to as hybrid methods. The predominant hybrid functional used by chemists is the B3LYP functional, which is written as a linear combination of HF exchange and local- and gradient-corrected exchange and correlation:

$$E_{xc}^{B3LYP} = aE_x^{HF} + (1-a)E_x^{Slater} + bE_x^{B88} + cE_c^{LYP} + (1-c)E_c^{VWN} \tag{27}$$

The weighting parameter a determines the extent of replacement of the Slater local exchange (E_x^{Slater}) by the exact HF exchange (E_x^{HF}); b control the addition of Becke's gradient-correction to the exchange functional (E_x^{B88}); c defines the inclusion weight of the LYP correlation (E_c^{LYP}) and the VWN correlation (E_c^{VWN}) functionals. (The difference between LYP and VWN is that the first is a gradient-corrected correlation functional, whereas the second is a local correlation functional.) The three coefficients were optimized by minimizing the average absolute deviation of theory from experiment for 116 atomic and molecular properties (56 atomization energies, 42 ionization potentials, 8 proton affinities and 10 first-row total atomic energies).

An important advantage of the DFT methods, when compared to wave-function-based methods, is the lower scaling. For DFT methods, the dependency of computational time (t) on the number of basis functions (N, which can also be approximately considered as the number of electrons) is $t \sim N^{\alpha}$ ($\alpha \approx 3$), while α is larger for wave-function methods. For example, in the case of HF methods, $\alpha = 4$. The relatively low computational cost makes the DFT methods possible to be applied to large systems. However, the employment of these methods is definitely affected by its accuracy, in particular the accuracy on geometry and energy.

Even though DFT seems to be an effective tool with several advantages, it is important to notice that it is not perfect and comes with several deficiencies, involving

mainly self-interaction errors, near-degeneracy errors and the lack of description of Van der Waals interactions. In wave-function methods, the artificial repulsion between an electron and itself is effectively canceled by an exchange term. In DFT, Coulombic terms are described exactly, but the exchange is described by an approximate functional. These terms do not cancel in DFT, leading to so-called self-interaction errors. This error artificially stabilizes delocalized transition states and tends to decrease the energy barrier heights.

The near-degeneracy error is due to the inherent description of the wave function in as a single determinant (the non-dynamical correlation is lacking). In contrast to the effect of the self-interaction error, this one tends to increase the height of the barrier. Therefore, there is a substantial cancellation effect between this two errors. In practice, the B3LYP functional appears to be built to balance these errors as well as possible. For high-electron-delocalization reactions, like hydrogen atom or proton transfer reactions, the barriers are usually underestimated since the self-interaction error prevails in these cases, while in the case of most other reactions it tends to be overestimated, since the near-degeneracy error predominates.

The third deficiency of DFT is the lack of a description of Van der Waals interactions. This deficiency often leads to exaggerated repulsion when the atoms are forced close to each other, which usually happens in systems with several large substituents or ligands.

2.3.3. Basis Sets

The quality of the results obtained when solving the Schödinger equation, be it using HF methods or DFT methods, depends on the basis sets we use to solve it. Basis sets are nothing more than a collection of known functions through the linear combination of which we can reproduce an unknown function. This functions usually take a shape similar to those of the atomic orbitals. As it would be expected, the more functions one uses, the better are the results expected to be. However, this brings about a major drawback: the greater the number of functions, the larger the time it takes to compute the result. This relationship between time and precision must be considered before attempting to make any calculation so that there is a good balance between both variables.

Basis sets can be divided between several different types, although those that are mostly used are the Slater-type orbitals (STOs) and Gaussian-type orbitals (GTOs).

STOs offer a good description for the electronic density around the atom. Nevertheless, it is worth noting that for systems with several atoms, where the atomic orbitals are centered in different atoms, the integrals that compose these functions are not solvable analytically when there are more than two centers. This limitation restricts the use of the STOs.

GTOs are the most commonly used type of basis sets. Contrarily to STOs, these GTOs can be solved analytically for multi-electron integrals, however this types of functionals do not provide a good description of the electronic density, particularly for distances very near or very far from the nucleus, which demands that we use a larger number of Gaussian functions. Despite these limitations, the use of GTOs is preferred to STOs because the higher number of functions used is balanced by their mathematical simplicity, which lowers the computational cost. The results obtained using GTOs can be further improved through the use of a contracted Gaussian basis set, which are generated through linear combination of uncontracted Gaussian functions (primitives) with fixed coefficients so that they form a smaller set of functions. Contracted GTOs enhance the description of orbitals with electrons closer to the nucleus, when compared to uncontracted GTOs, but they still do not correctly emulate the real behavior of these electrons. However, since the electrons closer to the nucleus do not usually play a major role in a chemical sense, this approximation is good enough to describe them without compromising the results.

Even though basis sets present several mathematical advantages, they lack in that they are not very flexible, i.e., they are not good at describing the deformation that occurs in orbitals when they are near other atoms. In order to give basis sets more flexibility, several methods, including polarization functions and diffuse basis functions, have been developed.

Polarization functions are nothing more than a set of Gaussian functions one unit higher in angular momentum than what are present in the ground state of the atom, which will increasing the flexibility of the basis set in the valence region in the molecule, since the electronic density can polarize the orbitals in energetically favorable directions.. This method is especially convenient when the distribution of electronic density around the nucleus is not isotropic.

For excited states and anions where the electronic density is more spread out over the molecule, some basis functions which themselves are more spread out are needed (i.e. GTOs with small exponents). These additional basis functions are called diffuse functions and they are normally added as single GTOs.

2.4. Hybrid Methods (QM/MM)

When studying the catalytic mechanism of enzyme, the use of computational methods has several advantages over experimental methods. They are faster and cheaper, and allow for the characterization of transition states and intermediaries, not only structurally but also energetically. However, when using computational methods, we must consider that enzymes are systems constituted by thousands of atoms, which make the resolution of the Schrödinger equation (or the calculation of the electronic density) considerably more complex. In order to determine the energy of such systems we can make use of very accurate methods, since that would make the calculation much more computationally demanding. Still, considering that when studying enzymatic mechanisms one must account for the forming and breaking of bonds, the method used must handle electrons explicitly.

In order to solve this problem, hybrid methods were created, which allow the system to be divided in several parts, each of which treated with a different theoretical level. If we were to divide the system in three parts, they could each be represented as follows:

- An internal layer, which includes all residues involved directly in the catalysis (active site) and the substrate. This layer should be treated with the highest theoretical level (for instance, post-Hartree-Fock or DFT), since it will in this layer that the change in chemical bonds will happen.
- An intermediate layer, which contains those residues closest to the active site. This layer emulates the short-range interactions between the active site and the rest of the enzyme, maintaining its structure and stabilizing it during the reaction. This layer should be treated with a medium level of theory, using for example a semi-empiric method.
- An outer layer, which includes the rest of the enzyme, emulating the proteic environment and is to be treated using the lowest level of theory, such as molecular mechanics.

Using different levels of theory for different layers will allow us to save time in the calculations and include a larger portion of the system. High levels of theory are only used for a restricted number of atoms and therefore the running time is shortened. However, since we also simulated the atoms that are surrounding the active site, the environment in which the reaction takes place resembles more closely that in which

reaction takes place in vivo. Overall it is expected that using this technique we will be able to gather results that are more precise in a smaller time frame.

Nowadays there are several different types of hybrid methods, but overall they can be divided in two categories: addictive and subtractive. This distinction is made by taking into account the way the final energy is calculated. In the case of addictive methods, the final value for the energy is calculated by adding all the individual values obtain for each layer plus a coupling term, which describes how the single layers interact with each other. For subtractive methods, the final energy calculation can be a bit more complex. We shall explore subtractive methods further in the next section by illustrating how the ONIOM method works.

2.5. ONIOM

As referred previously, ONIOM (Our own N-layered Integrated Orbital and Molecular mechanics) [64-67] is a hybrid subtractive method in which the system we want to study is divided in different layers, each of which treated with different levels of theory.

With the ONIOM method, systems are usually decomposed in two layers: the high layer, which is treated with a more precise mehtod (quantum mechanics) and the low layer, which is treated with a faster and less accurate methods, frequently molecular mechanics. The total energy of the model is calculated using the energies of each part calculated individually and according to the following expression:

$$E_{ONIOM} = E_{H+L}^{MM} - E_{H}^{MM} + E_{H}^{QM} \tag{28}$$

E_{H+L}^{MM} and E_{H}^{MM} refer to the energies of the whole system and the high layer, respectively, calculated using a low level method, and E_{H}^{QM} is the energy of the same high layer but this time calculated with a high level method.

One of the main problems with ONIOM (as well as other hybrid methods) when used for systems composed of a single macromolecule, which is then divided in different regions, relates to how the interface between the two regions should be handled. We must bear in mind that when separating both QM and MM subsystems we are breaking covalent bonds between the atoms, and therefore creating unpaired electrons in the high layer. An easy and elegant way of dealing with this is by introducing link atoms. In this metodology, the layers are separated from one another and at the broken covalent bonds link atoms are added, usually hydrogen atoms, which allow the calculation of a

meaningful electronic structure and therefore a reasonable energy value for the model system

Another problem as to do with how the electrostatic interaction between both layers should be calculated. This can be done using two different methods: electrostatic embedding or mechanical embedding. Electrostatic embedding includes the polarization of the QM region by the MM charge distribution in the QM calculation, which means that the electronic density/wavefunction of the high layer is modified and optimized according to the atoms on the low layer. Mechanical embedding, in the other hand, the electronic structure of the high layer is calculate in vacuum, without any influence from the low layer, and atomic partial charges are assigned to QM atoms based on the calculated electronic structure. Both approaches have their advantages and disavantages: electronic embedding yelds more accurate results, but is also more costly in computationally, so when choosing between one or the other one should bear this differences in mind.

CHAPTER 3

RECEPTOR-BASED VIRTUAL SCREENING PROTOCOL FOR DRUG DISCOVERY

In the pharmaceutical arena, virtual screening is normally regarded as the top CADD tool to screen large libraries of chemical structures and reduce them to a key set of likely drug candidates regarding a specific protein target. This chapter provides a comprehensive overview of the receptor-based virtual screening process and of its importance in the present drug discovery and development paradigm. Following a focused contextualization on the subject, the main stages of a virtual screening campaign, including its strengths and limitations, are the subject of particular attention in this review. In all of these stages special consideration will be given to practical issues that are normally the Achilles heel of the virtual screening process.

Adapted from reference [68]

In this review, Diana Gesto wrote several parts of the original manuscript, with special focus on section 3.4 (and all its subsections), helped with its editing and revision.

3.1. Introduction

The process of drug discovery is very complex and requires an interdisciplinary effort to design effective and commercially feasible drugs. The objective of drug design is to find a drug that can interact with a specific drug target and modify its activity. The drug targets are generally proteins that perform most of the tasks needed to keep cells alive. Drugs are small molecules that bind to a specific region of a protein and can turn it on or off. Some very powerful drugs, such as antibiotics or anticancer drugs, are used to completely disable a critical protein in the cell. These drugs can kill bacteria or cancer cells.

It is generally recognized that drug discovery and development are very time and resource-consuming processes and the whole process is often compared to searching for a needle in a haystack. It is estimated that a typical drug discovery cycle, from lead identification to clinical trials, can take 17 years with a cost of 800 million US dollars. In this process it is estimated that five out of 40,000 compounds tested in animals

eventually reach human testing and only one in five compounds that enter clinical studies is approved. This represents an enormous investment in terms of time, money and human resources. It includes chemical synthesis, purchase, and biological screening of hundreds of thousands of compounds to identify hits followed by their optimization to generate leads, which require further synthesis. In addition, predictability of animal studies in terms of both efficacy and toxicity is frequently suboptimal. Therefore, new approaches are needed to facilitate, expedite and streamline drug discovery and development, save time, money and resources.

On October 5, 1981, Fortune magazine published a cover article entitled “Next Industrial Revolution: Designing Drugs by Computer at Merck”. Some have credited this as being the start of intense interest in computer-aided drug design (CADD)[69].

CADD is defined by the IUPAC as all computer assisted techniques used to discover, design and optimize compounds with desired structure and properties. CADD has emerged from recent advances in computational chemistry and computer technology, and promises to revolutionize the design of functional molecules. The ultimate goal of CADD is to virtually screen a large database of compounds to generate a set of hit compounds (active drug candidates), lead compounds (most likely candidates for further evaluation), or optimize known lead compounds, *i.e*. transform biologically active compounds into suitable drugs by improving their physicochemical, pharmaceutical and ADMET/PK (pharmacokinetic) properties[70].

The fast expansion and popularity of this field of research has been made possible partially by the advances in software and hardware, computational power and sophistication. On the other hand, the knowledge of the 3D shapes of proteins, nucleic acids, and complex assemblies are fundamental to understand all aspects of potential drug targets. It is remarkable that, from 1970 to 2004, 50,000 structures have been deposited on the protein databank, in 2014 this number has tripled to 150,000 and in 2018 it is expected that this latter number doubles. In addition, the increasing digital repositories containing detailed information on potential drugs and other useful compounds provide goldmines for the design of new drugs.

CADD is widely used in the pharmaceutical industry to improve the efficiency of the drug discovery and development pipeline. One method that was quickly adopted was the virtual screening of large compound databases against drug targets. The goal is to select a set of molecules with desirable properties (active, drug-like, lead-like) targeting a specific protein and eliminate compounds with undesirable properties

(inactive, reactive, toxic, poor ADMET/PK). The computational methodologies used for this purpose are known as virtual screening methodologies.

The generic definition of virtual screening encompasses many different methodologies, which are generally divided in two main classes: the ligand-based virtual screening methods and the receptor-based virtual screening methods.

Ligand-based virtual screening methods aim to identify molecules sharing common features, both at the chemical and physical levels grounded in the assumption that similar compounds can have similar effects on a drug target[71]. These methods normally discard all information related to the drug target and focus exclusively on the ligand. Within the lock-and-key paradigm, these approaches compare different keys, and neglect the lock. Thus, the model of the receptor is only implicitly built based on what binds to it[72]. The main downside of these methods is that substantial activity data regarding the compounds that are studied are required to get reasonable results.

Receptor-based virtual screening methods, also called structure-based methods, require the existence of a 3D structure of the target. These methods involve explicit molecular docking of each ligand into the binding site of the target, producing a predicted binding mode for each database compound, together with a measure of the quality of the fit of the compound in the target-binding site. This information is then used to sort out ligands that bind strongly to the target protein from ligands that do not. Receptor-based approaches are gaining considerable importance over ligand-based techniques, particularly as more and more 3D structures of target proteins are determined and become available, and also because the results tend to be more reliable and accurate. The current state-of-the-art of receptor-based virtual screening is reviewed in this chapter, and general approaches, successes and pitfalls associated with the technology are highlighted.

3.2. The Screening Process

Receptor-based virtual screening encompasses a variety of sequential computational stages, including target and database preparation, docking and post-docking analysis, and prioritization of compounds for experimental testing. A typical workflow of a receptor-based virtual screening is presented in Figure 17. All stages of this workflow depend on sound implementation of a wide range of computational techniques that will be discussed in detail in the following sections. In each section special attention will be given to practical issues that are normally the Achilles heel of

the virtual screening process. Since in this book chapter only the receptor-based virtual screening will be reviewed, we are going to adopt the general term, virtual screening (VS), to describe this type of screening methodologies.

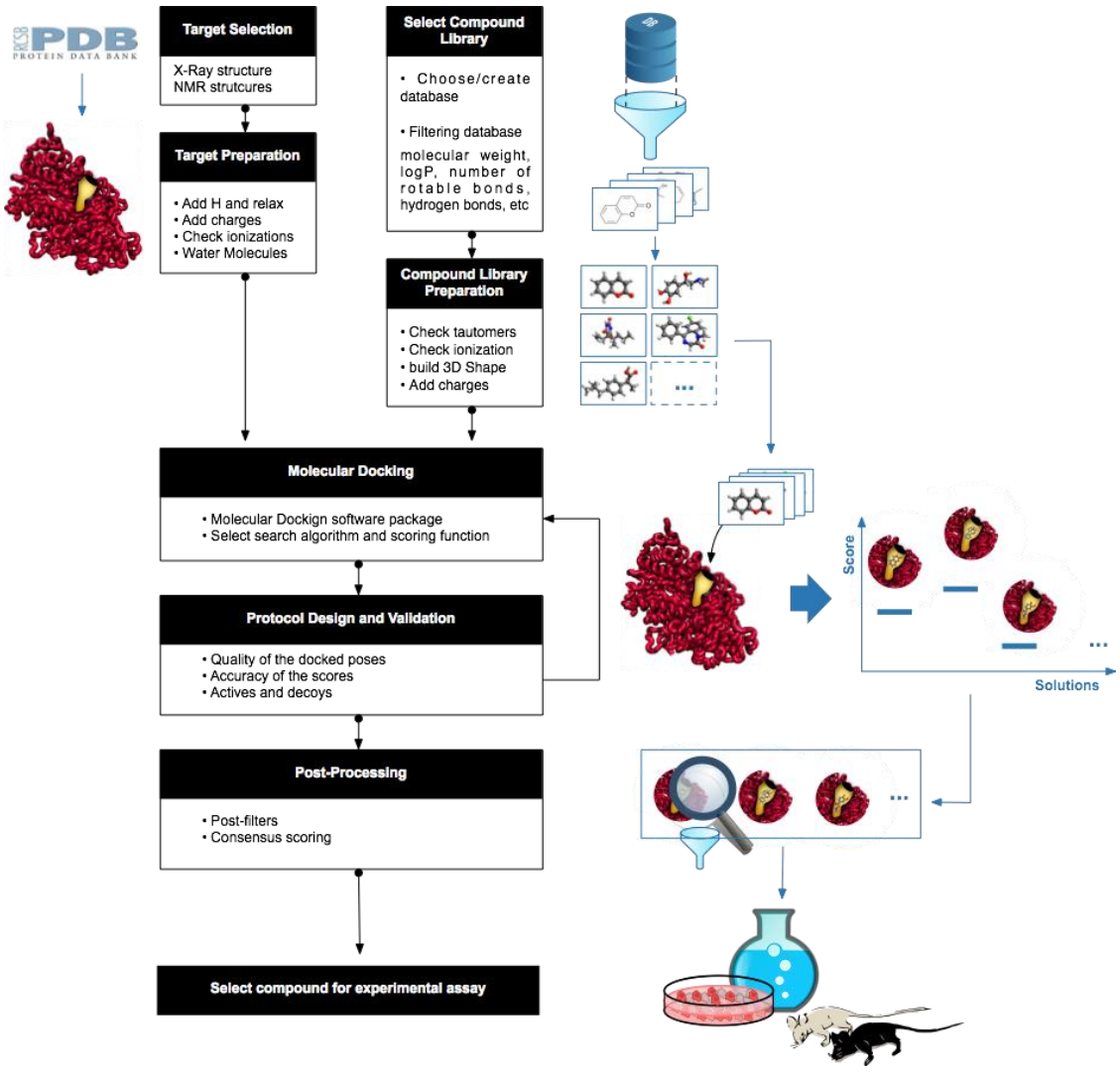


Figure 17 - General workflow of a receptor-based virtual screening. The typical workflow consists of a preparation phase for the database and the target, followed by a molecular docking phase, and concludes with the post-processing and compound selection phases.

3.3. Target Selection

Target selection is among the first stages of a virtual screening campaign and it is pivotal for a successful drug development process. Among the four types of macromolecules that can be targeted (proteins, polysaccharides, lipids and nucleic acids) with small-molecule compounds, proteins, and within those enzymes, are generally the first choice, since their binding pocket properties allow for high specificity,

potency and low toxicity. When considering a potential protein target to modify a disease it should be pondered if it is advantageous to select an upstream, widely implicated target or, instead, a downstream target, very specific to the pathway that we want to tackle.

Once a protein target with the potential to modify the disease has been identified, it is time to get its 3D structure. The Protein Data Bank is the leading repository for experimentally determined 3D structures of large biological molecules. This database is therefore the first approach to retrieve a protein 3D structure for a VS campaign. In the case that the experimental 3D structure of the protein does not exist, then homology-modeling methods can be used to build it. There are several examples in the literature showing that these homology models can be used with success in VS campaigns[73-75].

3.3.1. Binding Site Detection

Once the 3D structure of a protein has been obtained then it is possible to evaluate its *druggability* score. The *Druggability* can be understood as the capability that a receptor has to bind molecules with drug like properties. This depends of course on the ability of the molecule to favorably interact with a particular pocket or cleft in that protein. The location of these binding sites is easy when a ligand has been co-crystallized explicitly with the target protein. However, when this sort of information is not available, the location of the binding site can be cumbersome. In these cases computational tools can be used to identify and characterize potential binding sites. Among the available computational tools, some algorithms rely mostly on geometric characteristics to search for binding pockets, such as POCKET[76], LIGSITE[77], SURFNET[78], SPHGEN[79], FPOCKET[80], etc., while others, such as Q-SITEFINDER[81], GRID[82-83], POCKETPICKER[84], FLAPSITE[85] , CS-MAP algorithm[86], in order to calculate the energy of probes interacting with potential binding sites to identify and rank them. Geometry-based algorithms are usually prized because they are fast and robust in dealing with structural variations or missing atoms/residues in the input structure[87]. Energy-based algorithms, on the other hand, are often more sensitive and specific[88]. Despite the distinctive approaches, the performance is very similar and both methods can correctly predict 95% of the known binding sites[89].

3.3.2. Target Preparation

After the target has been defined and the most druggable binding site chosen, it is necessary to prepare the target for docking. The general steps in target preparation

require removing solvent and ligand molecules, adding hydrogen atoms, setting up bond orders and formal charges, capping chain termini and defining amino acid protonation states (atom types). It might also be necessary to refine the crystallographic structure and define the binding site portions that will be left flexible. Target preparation is usually overlooked but the effect on virtual screening enrichment might be considerable.

3.3.2.1. Structure Refinement

Crystallographic structures are pledged with uncertainty. A common example is the identification of the atoms in amide groups and in imidazole rings that are difficult to distinguish. For example, 180° flips of the terminal chi angle still fit in the electron density. These flips, in addition to the protonation state of aspartate, glutamate, lysine and histidine residues, the two tautomers of histidine residues or hydrogen orientation in hydroxyl and thiol groups can alter significantly the hydrogen bond network of the binding site. Sastry G. et al.[90] have shown that careful attention to these details could improve virtual screening performance on 20 out of 36 targets.

Another type of structure refinement that has a positive impact in the performance of the VS campaign is structure relaxation. A simple structure minimization removes clashes and tensions and improves the hydrogen bond network, two conditions that can be important to improve ligand fitting in the binding pocket of the target protein.[90]

3.3.2.2. Water Molecules

The 3D structure of proteins are often populated with a series of water molecules located inside the binding pocket. Some of these water molecules establish important interactions between the protein and the ligand and are thus important to take into account when studying protein-ligand interactions[91]. However, in many cases the water molecules are not important and may affect negatively the VS campaign, since they can occupy some region on the binding pocket that is required for the binding process. There is no perfect strategy to determine which active-site waters are important for ligand binding and which are not[92]. Therefore it is always good practice to analyze with caution the role of the water molecules inside the active site. To this end, several models of the protein with a different number of water molecules inside the active site should be tested and the results should be confronted with experimental data whenever it is available.

3.3.2.3. Metals

The preparation of metal sites in the receptor should be performed in accordance with the requirements of molecular docking programs. Scoring functions use various terms and formalisms to describe the interaction between docked ligands and the metal atom. In the simplest cases, scoring functions reward the proximity of potentially coordinating atoms, such as oxygen and nitrogen, to metals in receptor molecules. Other simplistic approaches treat metal atoms as regular hydrogen bond donors, including metal coordination in the hydrogen-bonds energy. In these situations, special attention is not required in the receptor model, but the protonation state of docked ligands might be important.

Some specialized approaches attempt to improve the docking accuracy and binding affinity prediction by re-parametrization of the metal ion force field, which often results in significant improvements[93-94]. For instance, the metal-ligand interaction is still modeled by traditional electrostatic formalisms, but fine tuning of the electrostatic interactions by polarizing ligand charges using the metal ion and its protein ligands in a QM/MM calculation should be used instead.

3.4. Ligand Selection

With virtual screening we have the possibility to quickly test a great number of compounds without much effort. However, it is obviously impossible to screen the entire chemical space for a single target in a timely manner, which means that we must somehow restrict the number of compounds to be tested. In order to achieve a manageable library we need to filter out some molecules in advance, which can be achieved with the help of different methods. Still, it is important to bear in mind that for a successful virtual screening campaign, it is desirable to have a database with as much variety as possible.

3.4.1. Databases

Before starting a new virtual screening campaign, it is necessary to collect all the structures that we want to test. If the search space is very limited, and we know which kind of molecules is likely to bind our target, we can draw our own structures and readily start the docking process. However, for most campaigns this is not the case. Usually,

before docking, it is necessary to build a library of potential ligands, which can have thousands of structures that will eventually be tested.

In recent years several databases of chemical structures have been developed, which can be easily accessed by most people who wish to build a compound library. These databases not only store the structure of these molecules, but also many chemical and biologically relevant information[95].

One of the most commonly used compound database is ZINC[96]. ZINC is a free database with over 35 million purchasable structures, including more than 4.5 million clean leads. The structures can be obtained in different formats, and several structural/functional filters are available to further limit the search space. Other databases include ChemSpider (32 million+ compounds), ChemDB[97] (5 million compounds) and PubChem[98] (63 million unique structures). Furthermore, all major pharmaceutical companies also have in-house corporate libraries, comprising several million compounds.

3.4.2. Reducing the Search Space

In theory, any molecule can be a ligand to some target. Nevertheless, for a compound to be considered a drug it needs other characteristics that make it suitable to be administered to the patient. For instance, if a ligand is known to have toxic, mutagenic or teratogenic properties, it can be automatically excluded as well as inorganic, insoluble, reactive and aggregating molecules. Also, since the ultimate goal of virtual screening is the creation of a new drug, we can easily reduce our search space if we focus solely on compounds that can be synthesized. There is no point in testing molecules that cannot exist outside the virtual world.

In order to evaluate if a new substance is a good candidate to become a drug or not, scientists have come up with the concept of druglikeness. This means that in order to be considered drug-like, a new molecule must have some of the characteristics shared by the majority of drugs known so far, which are related to the bioavailability of the compound after administration.

Currently, there are several methods that can be used to evaluate how drug-like a certain substance is, which allow us to reduce our database. Among them we have simple counting methods and functional group filters.

It is important to point out that this exclusion method is not infallible, as there can be good drug candidates that do not abide by these rules. It is also worth mentioning that

even if a compound follows all these rules, it is not guaranteed that it will pass all stages of the clinical trials and, in fact, it can be excluded in the earlier stages.

3.4.2.1. Counting Methods

Counting methods allow us to easily limit the search space by selecting only molecules that have certain properties, which are associated with other drug-molecules. Counting methods take into account characteristics like the partition coefficient (log P), molecular weight, and hydrogen bonding groups, all of which are related to bioavailability. It is expectable that if we include only compounds that present favorable characteristics our chances of finding a new drug that successfully passes the clinical trials will increase immensely.

One of the most famous counting methods is known as the Rule of Five and was developed by Christopher A. Lipinski[99]. It presents a rule of thumb that can be used to evaluate the drug likeliness of a compound, and dictates the following:

- The number of hydrogen bond donors must be 5 or less;
- The number of hydrogen bond acceptors must be 10 or less;
- Less than 500 Da of molecular weight;
- Log P smaller than 5.

The name Rule of Five originates in the fact that all numbers are multiple of 5. Although there seems to be quite a few cases to which these rules do not appear to apply, most of them are antibiotics, vitamins, antifungals, and cardiac glycosides.

Further work done by Ghose et al. allowed for the improvement of the Rule of Five. After analyzing the physicochemical property profiles of more than 6000 drugs (included in the Comprehensive Medical Chemistry Database), they established that 80% of these had the following properties[100]:

- Log P between -0.4 and 5.6;
- Molecular refractivity between 40 and 130;
- Molecular weight between 160 and 480;
- Number of atoms between 20 and 70;
- Polar surface area smaller than 140 Å2.

Veber et al. proposed further that the number of rotatable bonds should be smaller than 7, in order to enhance oral bioavailability[101].

3.4.2.2. Functional Group Filters

Functional group filters are based on the fact that certain functional groups are not suitable in a drug, either because they are highly reactive or toxic for the organism. Functional groups known to cause damage to the organism, such as potentially mutagenic or teratogenic groups are often discarded, since the compounds they form are usually also harmful. On the other hand, reactive groups (alkyl-bromides, metals, etc.) can often originate hits, which are subsequently not pursued because most of the times they give rise to false positives. Moreover, specific reactive groups such as alkyl-bromides, metals, etc., often give rise to false positives. This allows us to save time by focusing solely on compounds that are more likely to be good leads.

3.4.3. How to use Filters

Usually we do not want to use only one type of filter. While filtering out compounds based solely on one aspect will definitely reduce our library, it will still be rather large and contain molecules that, had we used a different filter, would not be there. Since our ultimate goal is to find the most leads in the least amount of time it is good practice to use more than one type of filter in order to enrich our library with the molecules that have the best odds of being a hit.

Another method for further reducing the search space is to cluster molecules based on their similarity. Similarity between compounds can be calculated with different methods, the most often used being the Tanimoto coefficient (section 3.4.3.2.).

3.4.3.1. Similarity Searching

When doing virtual screening, one of the most important parts is to find as much information on the target as possible: to know which the most favorable region to bind an inhibitor is, where the active site is, what kind of molecules is more likely to become hit compounds. However, we sometimes have little more knowledge about the target than its 3D structure and have to try and make the best out of it. In cases where

information is scarce, in order to maximize our chances of finding an active hit, it is best to use libraries with as much diversity as possible.

The problem with libraries built on diversity is that they can have a very larger number of compounds, which might make them impossible to screen in a reasonable time frame. We can nonetheless restrict the search space if we assume that similar molecules usually display similar biological activities. What this means is that if we group molecules in sets based on their identity and use only one representative from each set we can have a highly diverse library with fewer elements, thus eliminating chemical redundancy. Although this assumption is widely accepted, one can easily find several arguments that contradict it as, sometimes, subtle changes in a molecule can cause its activity to change sharply, or the reverse, in which a structure is altered significantly and yet its activity remains identical. Still, this method of restricting the search space has been shown to work and it has been proved that a set of compounds related to an active hit have more biologically active molecules than a randomly generated set. However, if the chemical space is overly restricted, there is a reduced probability of finding hits, because even similar molecules can have different activities. We must therefore select a similarity cutoff big enough to broadly restrict the library size, and yet small enough so that all the molecules in a given set can be well represented by a single compound.

Similarity search is not used only in cases where we have little to no information about the target. It can also be used when we already know some of the characteristics that an active compound must have (for example if we have information on a hit) and wish to find other molecules with higher affinity. In this case, we build our dataset by gathering molecules that exhibit a certain degree of similarity to our substructure.

3.4.3.2. Tanimoto Coefficient

Similarity can be defined in several different ways. In chemistry (and especially in virtual screening) one of the most important indexes used to represent similarity is the Tanimoto coefficient (or Jaccard coefficient). The Tanimoto coefficient is a measure to determine how similar (or how distant) two objects are.

The Tanimoto coefficient is calculated by enumerating a list of attributes, which may or may not be present in a given object. To calculate how similar two different objects are, one simply needs to consider if these attributes are present in both objects or only in one of them. The coefficient that one gets by dividing the number of common attributes by the total number of attributes is the Tanimoto coefficient. In mathematical

terms, if we have two objects A and B, their Tanimoto coefficient will be given by the following expression:

$$T_c(A,B) = \frac{|A \cap B|}{|A \cup B|} \tag{29}$$

Another way to express this coefficient is if we consider the following table, which represents the presence (1) or absence (0) of features in two objects:

		Object B	
		0	**1**
Object A	**1**	a	c
	0	d	b

a – number of features present in object A and not in object B
b – number of features present in object B and not in object A
c – number of features present in both objects
d – number of features absent in both objects

$$T_c(A,B) = \frac{c}{a+b+c} \tag{30}$$

Based on this equation, it is easy to see that the Tanimoto coefficient can vary from 0 to 1, 1 meaning that both objects are identical (at least based on the characteristics we chose) and 0 meaning that they are completely different. To exemplify, let us consider ten different attributes for objects A and B, which are either present or not in the following manner:

A	0	1	1	0	0	1	0	1	0	0
B	1	0	0	0	1	1	1	1	1	0

If we create a table identical to the one above, we get the following:

		Object B	
		0	**1**
Object A	**1**	2	2
	0	2	4

To calculate the Tanimoto coefficient we must divide 2 (number of attributes present in both objects) by 8 (number of attributes present in either both or only one object), which gives us a similarity of 0.25. It is important to note that the Tanimoto coefficient does not take into account features that are absent from both objects, even when we are comparing more than two, while other methods, like the Euclidian coefficient, do. This indicates that different methods may give different results, depending on the number of features we consider.

The Tanimoto similarity is used in Chemistry to compare different molecules and substructures and to find out how identical they are in relation to one another. Similarity is calculated by selecting different attributes and checking whether they are present or absent in each compound.

3.5. Molecular Docking

Once the target protein and a database of compounds have been selected, molecular docking is ready to run. This is the stage that requires more computational cost and time in the VS and for this reason it is regarded as the heart of any virtual screening campaign.

Molecular Docking is a computational method that allows predicting the preferred pose and conformation of one molecule (ligand) in relation to a second one (often larger and called receptor), when the binding between the two forms a stable complex. The preferred orientation of the molecule in relation to the receptor can then be used to predict the strength of association or the binding affinity between both receptor and ligand. In a virtual screening campaign, the molecular docking stage is repeated, as many times as the number of compounds that exists in the compound library.

Currently, there is a relatively large and ever increasing number of molecular docking programs that can be used in virtual screening campaigns. Generally speaking, these programs have similar implementations. All of them involve the search for the preferred poses/conformations of the ligand in relation to the receptor. These

poses/conformations can be accomplished using two types of algorithms: the search algorithm and the scoring function. The search algorithm generates the various possible poses (conformations and orientations) to fit the ligand into the binding pocket of the receptor. The scoring function ranks the different poses and locations of the ligand that are generated by the search algorithm, and orders them by a score. This value should ideally represent the thermodynamics of interaction of the protein–ligand system in order to distinguish the true binding modes from all the others that are explored.

In the end of this process, the best-scored solution should correspond to a true binding conformation and should be close to the one that is observed experimentally, if such information exists.

The majority of molecular docking programs are developed to be fast since they are supposed to be applied to large databases of compounds. To this end, several assumptions and simplifications are included in the search algorithms and scoring functions that will be described briefly in the next sections.

3.5.1. Search Algorithms

The goal of a search algorithm is to provide enough degrees of freedom to the system, so that it will more accurately predict and/or reproduce experimental data (structures obtained from X-ray crystallography or NMR).

A real biological binary protein-ligand system requires searching over a space of (N*M + 6) dimensions (N and M parameters describe the protein and ligand 3D structures, respectively, and 6 are the rotation and translation components of the spatial arrangement of one unit onto the other). This high dimensional space is computationally untreatable, and to overcome this issue, the docking algorithms integrate different approximations to efficiently sample the search space. The number of available search algorithms is continuously increasing to optimize the speed, reliability and accuracy in sampling the relevant conformational space. The speed is particularly crucial for virtual screening campaigns, in which millions of different ligands are evaluated. Currently, the docking algorithms can be categorized into three main classes: rigid-body, flexible-ligand and flexible target methods.

Rigid-body Search Algorithms are the most basic and fastest algorithms to sample the conformational space in molecular docking. They do not take into account the flexibility of neither ligand nor receptor and therefore the final results are just based on the geometrical complementarity between both molecules. This simplest approach was

widely applied in the earlier protein-ligand docking studies and, currently, in protein–protein docking protocols or in the initial stages of virtual screening studies[102 103-104]. Popular implementations of this algorithm are found in ZDOCK[105], FTDOCK[106], SYSDOC[107], EUDOC[108], DOCK[102, 109], MSDOCK[110], LigandFit[111] and Glide[112].

Flexible-Ligand Search Algorithms consider the conformational space of the ligand while the receptor is kept rigid. Nowadays, these algorithms are the most popular ones and used in a wide range of molecular docking packages[113-115]. Taking into account the type of algorithm that is used to flexibilize the ligand they are divided in two main classes: systematic algorithms and random or stochastic algorithms.

Systematic search algorithms try to explore all the degrees of freedom of the ligand through conformational searches, fragmentation base methods or resorting to databases where this sort of information can be found. The systematic conformational approach is implemented in DOCK[109], the fragmentation base methods in LUDI[116], FlexX[117], DOCK[109], ADAM[118], Hammerhead[119], Surflex[120-121], eHiTS[122], FLOG[123] and the databases approaches in FLOG[123].

The random search algorithms sample the ligands' (or a population of ligands) conformational space by doing stochastic modifications in its conformation, which could be accepted or rejected based on a predefined probability function. This is generally implemented using Monte Carlo methods, Genetic Algorithms or Tabu Search methods. Monte Carlo (MC) methods take into account a Boltzmann probability function as the acceptance criterion for a newly generated ligand pose. MC algorithms dock the ligand inside the protein binding site through many random translations, rotations and conformations, decreasing the probability of becoming trapped in local minima. Prodock[124], ICM[125], MCDOCK[126], DockVision[127] and QXP[128] are some examples of docking programs that have an MC-based algorithm.

Genetic algorithms (GA) are a global searching strategy that belongs to the evolutionary programming methods with the purpose of finding solutions for search problems and, in the molecular docking case, of trying to find the pose closest to the global energy minimum for a given protein conformation. GA methods are heuristic algorithms that emerged from genetics and the theory of biological evolution. GOLD[129-130], AutoDock[131], DIVALI[132], and DARWIN[133] are some examples of docking programs that implement genetic algorithms in their search engines.

Tabu Search methods are meta-heuristic methods characterized by an iterative procedure that moves the ligand from one pose to another and imposes several restrictions to prevent revisiting previously considered poses. Earlier visited poses are

stored in a *tabu list* and the root-mean-square deviation (RMSd) of a new conformation, in relation to the previous ones, is calculated and used as criterion to accept or reject the new conformation relatively to the previous ones. PRO_LEADS[134-135] is the most popular docking program that uses a tabu-search algorithm.

Several protein-ligand docking studies have shown that the application of the flexible-ligand docking algorithms only give successful results when the protein is rather rigid and its 3D structure is representative of the target conformation in the docked complex. However, many proteins display significant structural changes upon ligand binding, such as the local rearrangement of side chains near the binding site. This is particularly evident for enzymes because they can possess different conformations to recognize their substrates and also for the transition state's stabilization along catalysis. This means that these small movements can have a high impact on the molecular docking results and the generation of false positive binding solutions[136].

To solve this challenging docking problem, some specialized search algorithms and computational strategies were developed to accurately account for the partial protein flexibility, in addition to the ligand flexibility, i.e. Flexible-Ligand and Receptor Search Algorithms. Nowadays, several docking programs offer this treatment. Approaches addressing the target flexibility can be classified as Molecular Dynamics (MD) and Monte Carlo (MC) methods, rotamer libraries, protein ensemble grids and soft-receptor modeling.

MD simulations and MC methods generate different configurations for the system, and their main advantage is that they are very accurate and can model explicitly all the degrees of freedom of the protein-ligand system and may also include the solvent if necessary. However, the high-dimensionality of the search space involved in these simulations tools, makes an ergodic exploration of the protein conformations unfeasible, due to the higher computational time required[137]. Rotamer library based methods are currently the most popular methods as they can represent the protein conformational space as a set of experimentally observed and preferred rotameric states for each residue side chain (Figure 18)[124, 138-140]. An ensemble of protein conformations, obtained from X-ray crystallography, NMR or MD/MC simulations, can be used as another strategy to include protein flexibility[141-142]. However, these search methods have two disadvantages: how the initial protein conformations are generated and how they are combined among themselves. A popular ensemble docking method is FlexE[143], an extension of the docking tool FlexX. The soft-receptor modeling method combines the information derived from several different experimental and computational protein conformations to generate one energy weighted average grid, which is subsequently

used to dock the ligands[141, 144]. This protein flexibility docking technique is the less computationally demanding approach, however, it cannot manage large-scale motions.

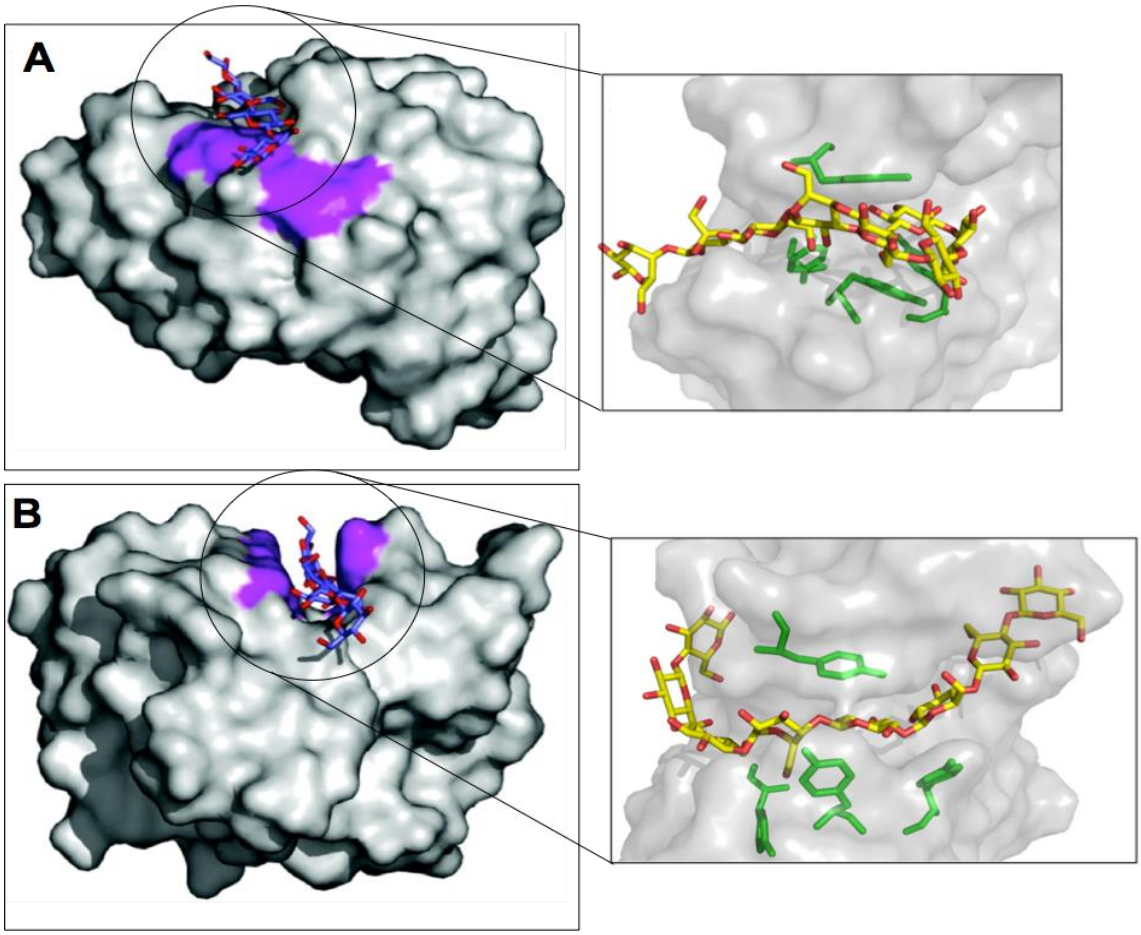


Figure 18 - Prediction of the binding pose of a carbohydrate into a carbohydrate binding module (Cbm) using two different molecular docking approaches. A: Flexible-ligand molecular docking protocol (Autodock[145]). B: Flexible ligand and flexible receptor molecular docking protocol (MADAMM[136]). The results have shown that in this case, it was very important to introduce some degree of flexibility into the binding site of the Cbm during the molecular docking stage. This created a suitable cleft into the Cbm structure that allowed an unbiased binding of the carbohydrate in it. The complexes generate by this process are in agreement with the available experimental data[146] and allowed to gather a better understanding of these proteins that are attached to enzymes that can decompose cellulose into glucose units[147].

As in a virtual screening campaign, hundreds or thousands of ligands are evaluated, the Flexible-Ligand and Receptor Search Algorithms are often discarded due to the time and computational cost that they require. Currently, the Flexible-Ligand Search Algorithms are used instead since they provide very good results.

3.5.2. Scoring Functions

The main goal of a scoring function is to calculate computationally an energy value that estimates the binding affinity between the protein and the ligand. There are a wide variety of different algorithms for predicting the binding free energy of a protein-ligand complex. These techniques differ significantly in accuracy and speed. If one wants to predict the binding free energy of only one ligand, very accurate but time-consuming techniques such as Free Energy Perturbation (FEP) or Thermodynamic Integration (TI) can be used. If the aim is, however, to compare binding free energies of hundreds or thousands of protein-ligand complexes as generated by virtual screening, then "scoring functions" are used instead.

A "scoring function" is a method that relies on several assumptions and simplifications. These methods are very fast since the complexity and computational cost required for the calculation of protein-ligand binding is dramatically reduced. However, the accuracy of the final results can be compromised, as a number of physical phenomena that determine molecular recognition are not included in the calculation or are modeled by predefined parameters that are obtained from experimental observations or quantum chemical calculations. The development of scoring functions is thus not an easy task, as it can have a major impact on the quality of molecular docking results.

Generally speaking, the accuracy of a scoring function can be evaluated taking into account its capability to follow the following criteria: i) it must be capable of estimating the interaction between the receptor and the ligand and this value should be close to the free energy of binding. ii) different binding poses must be ranked correctly, i.e. those that resemble most closely the experimental structures should be the best scored. iii) if multiple ligands are docked, their binding free energies need to be ranked accurately and it must be possible to discriminate between molecules that bind the target and molecules that do not. iv) a scoring function must be sufficiently fast to be applied in a docking algorithm[148].

Currently, the number of scoring functions available to assess and rationalize ligand protein interactions is large and increasing. Many algorithms share common methodologies with novel extensions, and the diversity in both their complexity and computational speed provides a plethora of techniques to tackle modern structure-based drug design problems. Roughly speaking, the scoring functions can be grouped into three main categories: force field scoring functions, empirical scoring functions and knowledge-based potentials.

Force-field-based scoring functions have been used for more than 2 decades and apply classical molecular mechanics energy functions to compute the binding energy between the receptor and the ligand and, in some cases, the internal energy of the ligand. Popular implementations of these scoring functions are D-Score that is based on the Tripos force field[149], and DOCK[150] and AutoDock[145, 151], both based on the Amber force field[152].

The functional form of the empirical scoring functions is often simpler than the force-field based scoring functions, although many of the individual contributing terms have counterparts in the force-field molecular mechanics terms. Currently, several empirical scoring functions are available in diverse molecular docking software, such as FlexX[117], F-Score[153], the Piecewise Linear Potential (PLP)[154], Chemscore[155-156], Glide SP/XP[157], SCORE[158], Fresno[159] and X-SCORE[160].

Knowledge-based scoring functions are pure statistical methods that are designed to reproduce experimental structures rather than reproduce binding affinities (such as the force field and empirical based scoring functions). These scoring functions use simple statistical potentials that estimate the frequency of occurrence or non-occurrence (*i.e.* negative data) of different atom–atom pair contacts and other typical interactions that are obtained from the structural information embedded in experimentally determined atomic structures. In this process, it is assumed that if an interatomic distance occurs more often than some average value, it should represent a favourable contact and vice-versa. In addition, the observed distribution of distances between pairs of different atom types must reflect their interaction energies. Muegges's Potential of Mean Force (PMF)[161-163], DrugScore[164-165] and SMall Molecule Growth (SMoG)[166] are the most popular examples of knowledge-based scoring functions.

3.6. Validation of the VS

Since many stages of the VS, as well as each of those stages, rely on many parameters, it is important to design a protocol to validate it. This will allow to have confidence on the computational results but also to decrease the number of false positives in the final results. Generally, the procedures employed to validate the VS campaign and in particular the molecular docking stage are classified in three groups: (1) quality of docked poses, (2) accuracy of scores or affinity estimates and (3) power to discriminate between active and non-active compounds.

3.6.1. Quality of the Docked Poses

The quality of the docked poses obtained from a molecular docking protocol is assessed by re-docking experiments. The term re-docking is used because the test consists in docking ligands for which the experimental binding modes have already been established, almost always by x-ray crystallography. The standard way to compare a docked pose with a known structure is to calculate the Root Mean Square deviation (RMSd) between the docked and experimental conformations of the ligand. It is common to exclude hydrogen atoms because they are rarely found in crystallographic structures. Cutoff values used to classify poses as correct or incorrect are usually around 2.0 Å.

RMSd values can be calculated in an adapted way, with the aim of compensating for symmetric conformations in the docked molecule[151, 167]. For example, a 180º rotation of a phenyl ring results in an equivalent conformation, but a standard RMSd calculation produces an artificially high value. Adapted methods search for the lowest RMSd within groups of exchangeable atoms, which are defined based on atom type or element.

An alternative metric to the traditional RMSd is to quantify how well the docked poses fit the experimental electron density of the crystallographic structure. It removes any bias of crystallographic models that may have been imposed when fitting the crystallographic model to the electron density, and it does not have any issue with symmetric conformations.

The re-docking test outcome is primarily related to the power of the docking engine and quality of the receptor model. However, it is worth to note that every docking engine must use a scoring function during the search to select among good and bad poses and finally lead to the correct binding mode. For this reason, re-docking tests are also related to the quality of the scoring function used in conjunction with the search algorithm.

When performing re-docking evaluations, it is advisable to perform visual inspection to double check low RMSd poses. Sometimes, the human eye can uncover issues not seen in the RMSd value.

3.6.2. Accuracy of the Scores

The process of scoring molecules takes place upon the prediction of binding poses. Correct poses influence the scoring outcome positively but sometimes inaccurate scores are generated for good quality poses. A direct assessment of the scoring function is important for the development and validation of a virtual screening protocol[168].

Experimental dissociation constants (Kd) or inhibition constants (Ki) constitute the most valuable information to support this validation procedure. These constants are directly associated with the free energy of binding, and are within the scope of physical properties modeled by a molecular docking program. On the other hand, IC50 values depend on experimental conditions, such as substrate concentration, rendering different experiments incomparable. Also, this kind of data can derive from experiments at the cellular level, where several additional factors influence binding in one way or another, making it difficult for molecular docking to reproduce the observed properties. Thus, IC50 values can only be useful if all data originates from the same experiment[169].

The scores produced for docked molecules should be indicative of the binding affinities in order to enable identification of active molecules in large and diverse compound datasets. These scores can be expressed in units of free energy or a number without physical meaning. Independently of the scoring units, better accuracy means an increased correlation with the binding affinity of the molecules. In this line of thought, a linear fit between scores and the experimental values allows one to predict the quality of the virtual screening, since a good fit means the relative affinity of a series of ligands is well reproduced. Typically, the ordinary least squares method is used, yielding a linear regression model where the r-free value is indicative of the quality of the fit. This statistical variable can be used to compare and choose among different scoring functions, and also to evaluate other parameters in the virtual screening setup[170].

3.6.3. Actives and Decoys

The most straightforward way to evaluate the performance of a VS campaign is to quantify its power to discriminate between active and inactive molecules. For this purpose, each molecule needs to be classified as active or decoy. It is rare to find explicit information on inactive molecules reported in scientific literature, even though some information is available in the Chembl and PubChem databases. Alternatively, random molecules are used as decoys, based on the assumption that they are not active for the target of interest. This has been most effectively introduced by the Directory of Useful Decoys approach[171-172].

In a perfect scenario, active molecules are scored significantly better than inactive ones, and the worst score for an active is better than the best score for an inactive. In reality, there is significant overlap of the distribution of scores for active and inactive molecules.

The statistical tools to quantify the performance of scoring methods given a ranked list of positive (active molecules) and negative (decoys) observations are well established. Most performance metrics are based on a cutoff or threshold value to classify molecules in active or inactive. Active molecules scored above the cutoff are called true positives, while decoys incorrectly classified as actives, are false positives[173]. Similarly, decoys under the cutoff score are true negatives and actives misclassified as inactive are false negatives[174].

Accuracy, precision and recall are typical metrics used to measure the performance of binary classification systems. In the realm of VS, enrichment factor (EF) is used instead, but it is conceptually similar to recall. Accuracy is the ratio between correctly classified molecules and the total number of molecules. Precision is the proportion of real actives (true positives) within all positives (true positives + false positives). Recall is the ratio between recovered actives (true positives) and the total number of actives (true positives + false negatives). Enrichment factor is defined as the ratio between recovered actives and the expected number of recovered actives using random scores. Note that all these measures are calculated upon discretization of the scores by a threshold.

The relative importance of precision and recall depends on the subsequent actions to be performed after the VS campaign. For example, if a small number of molecules is to be tested experimentally, high precision is critical to increase the odds of finding hits. On the other hand, in situations where a large number of compounds are selected from VS, lower precision can be afforded in order to allow for a greater recall, increasing the number of hits at the expense of a lower ratio of hits per inactive molecule. In order to account for a more complete picture of the performance, enrichment factors are usually reported for more than one threshold, which is represented as the score cutoff for which a determined portion of the chemical library or compound dataset. Typical enrichment factors are reported at around 1% of datasets, but up to 20% is not uncommon[175].

The choice of a threshold, even if tailored for a specific need, hinders a general view of the performance. The Receiver Operator Characteristic (ROC) curve was developed to graphically represent the overall performance of a ranking method, independently of a particular threshold. It consists in a plot of the true positive rate (TPR) in the y-axis and the false positive rate (FPR) in the x-axis. The TPR and FPR are the numbers of true positives and false positives expressed as a percent value of total number of actives and decoys, respectively. The area under the ROC curve (AU-ROC) is the probability of ranking actives better than inactives. A random ranking method corresponds to the equality line TPR = FPR, and has an AU-ROC of 0.5 (or 50%).

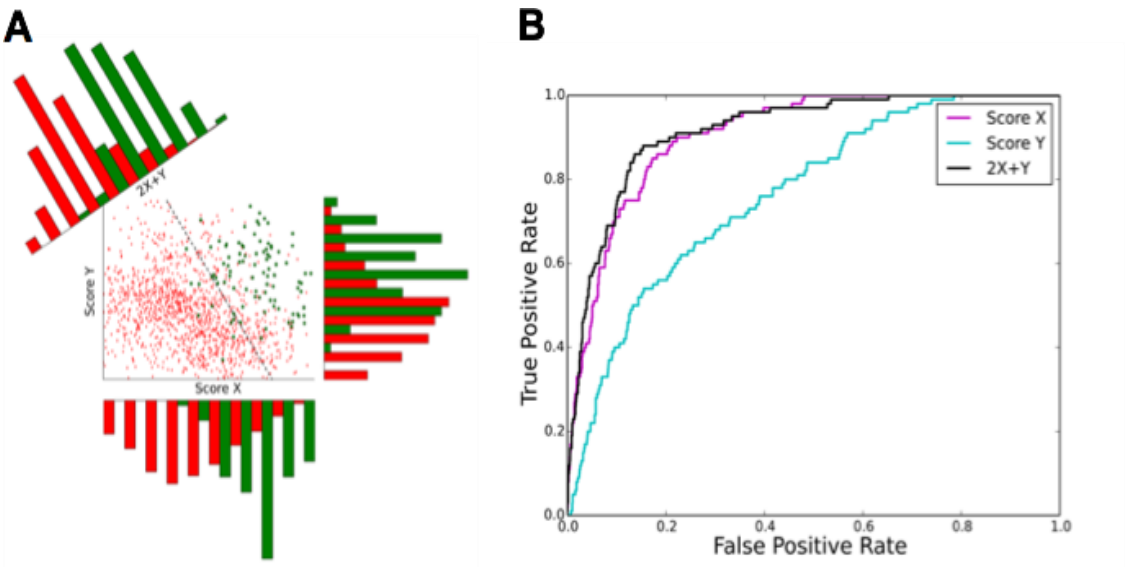


Figure 19 - Example of consensus scoring. A) the plot of score X versus score Y reveals that active compounds (green circles) are ranked with higher scores in both ranking methods. A linear combination of scores X and Y provides a better separation from decoy molecules (red dots), as it is illustrated by the black dashed line. The normalized distributions of scores X, Y and 2X + Y are provided by green and red bars for active and decoy molecules, respectively. B) The ROC curves associated with scores X, Y and the linear model 2X + Y show that score Y is the weakest of all scores. The consensus score has slightly better performance than score X.

In general, the number of compounds selected for further development in VS is significantly lower than the molecules in chemical libraries. Thus, it is more important to use these metrics for the top ranks of the dataset or, in other words, to evaluate the early recognition. Several tests have been developed for this purpose, generally as adaptations of the AU-ROC metric. An example of a simple adaptation is to calculate the area under a logarithmic plot of the ROC curve, which gives more emphasis to the performance in the early ranks. Another alternative is BEDROC, as it gives the possibility to control the earliness of the recognition. The performance of this kind of metrics is sensitive to the ratio between actives and decoys and its statistical power, or reliability, decreases if excessive earliness is requested[176]. As a rule of thumb, it is recommended to use AU-ROC for its good statistical power, together with one of the early recognition metrics (BEDROC) as well as enrichment factors at the relevant early ranks of the dataset.

Finally, an important aspect that is easily forgotten in this type of analysis is the scaffold variety. While using DUD (Directory of Useful Decoys)[172], it was observed that a single scaffold could be responsible for high enrichment. If such a situation occurs while validating a Virtual Screening protocol, the probability of finding new and different active molecules is severely limited.

3.7. Post-Processing Stage

Once all of the compounds from the library database have been docked into the binding pocket of the drug target, it is time to select which ones should carry on to the experimental testing. The easiest way of doing it, is to simply use the scores of the scoring function directly implemented in the docking algorithm, rank the compounds according to these values and take the top scorers for experimental testing. However this is not a straightforward process. At the end there are still too many compounds to test, or the compounds in the hit list resemble one another to such an extent that there is no point in testing all of them. Additionally, there are normally many undesirable false positives among the top scored list that need to be eliminated. To overcome these issues a selection of post-filters and /or consensus scoring methods are normally applied to limit the number of hits to be considered in subsequent drug development efforts.

3.7.1. Post-Filters

3.7.1.1. Visual Inspection

Visual inspection is naturally the number one option for post-filters. Such analysis helps to detect artifacts, such as incorrect metal coordination poses or badly oriented hydrogen bonds[177]. However this sort of analysis is not suitable for large-scale applications, when too many hits have been identified. It is normally advisable at this stage to apply automatic filters. A common strategy is to re-apply some of the filters initially used to reduce the number of species in the compound library, but now following more rigorous criteria.

3.7.1.2. Clustering Molecules

In many cases, at the end of a VS campaign, there are many compounds in the hit list that resemble one another to such an extent that there is no point in testing all of them. In these cases it is advisable to cluster the hit list into groups of similar compounds and select only a representative compound from each cluster to retrieve new and varied molecular scaffolds for further enhancement.

3.7.1.3. Consensus Scoring

It has been shown that the main reason for the misranking of compounds in VS lies in the scoring functions of the molecular docking stage that very often do not rank the correct conformational solution first, and more importantly, that often fail in the comparison of the binding affinity of different ligands. As a result, they fail to distinguish inactive from active compounds, therefore causing many false positives to be among the top scorers of a single ranking list.

Several groups suggested approaches to improve the selection of the correct bioactive conformation out of the set of generated conformations. One possible approach is the so-called “consensus scoring” in which docked poses are rescored with several different scoring functions. This method was introduced by Charifson et al[178]. In their study, they docked the ligands with their in-house docking tool Gambler, rescored the generated docked ligand conformations with 2-3 scoring functions and took then the intersection of the top N% of each of the sorted ranking lists (consensus list). They considered altogether thirteen different scoring functions in their study and have shown that the consensus lists contain significantly less false positives than obtained with a single scoring function, and conclude that a combination of scoring functions significantly enhances hit rates (40). This shows that the false positives of one scoring function are not the same as the false positives of another, resulting in a significant elimination of false positives when generating the consensus. A second study was published by Stahl et al, who carried out a detailed analysis using FlexX as the docking engine and four scoring functions (FlexX, PLP, PMF, Drugscore) for rescoring[179]. They mainly confirmed the result of Charifson et al, stating that consensus scoring is generally successful when two scoring functions are combined that perform well individually.

3.8. Future Developments and Perspectives

It is generally recognized that drug discovery and development are time and resources consuming processes. There is an ever-growing effort to apply computational power to the combined chemical and biological space in order to streamline drug discovery, design, development and optimization. In the pharmaceutical industry, CADD is being utilized to expedite and facilitate hit identification, hit-to-lead selection, optimize the absorption, distribution, metabolism, excretion and toxicity profile and avoid safety issues.

The contribution of computational methods, such as VS, to drug discovery is no longer a matter of dispute. All the world's major pharmaceutical and biotechnology companies use computational design tools. Nowadays, it is estimated that CADD accounts for 10% of pharmaceutical R&D expenditure and that they will rise to 20% by 2016[180].

There is a good number of successful studies where CADD, and in particular VS, aided in the development of new drugs. However, and in spite of the very positive picture that is often drawn with these methodologies, the effectiveness and impact of VS are currently limited by major scientific problems that are far from being solved. Firstly, the problems include imprecise pose scoring and binding energy predictions as well as incorrect similarity-based compound rankings all of which require a time consuming follow-up analysis to select candidate leads on the basis of knowledge or intuition. In addition, although ligands are commonly handled with full flexibility, the protein flexibility is still, at best, only partially considered. Further studies are still necessary to tackle this issue and address the induced-fit problem. Moreover, the dynamic inclusion of water molecules during the docking process, to take account of eventually important water-mediated hydrogen bond bridges between the ligand and the protein, could increase the efficiency of the approach.

In order to turn VS the key step in drug development, it is still necessary to prove that the aspects on which it depends on are correct and reproducible. This can be achieved either by scrupulous experimental validation, or by the development of new virtual screening methodologies. The interplay between computational modeling and experimental research is therefore a decisive stage where the inputs from each of these disciplines are essential for their mutual growth.

Despite these limitations, VS is still the best option available nowadays to explore a large chemical space in terms of cost effectiveness and commitment in time and material, as it allows access to a large number of possible ligands, most of them easily available for purchase and subsequent test. With the increasing number of targets identified by genomics and proteomics, and improved methodologies capable of predicting better hit rates and better predictions of geometries, VS methodologies will gather an even more preponderant role in drug design in the near future.

CHAPTER 4

CHOLESTEROL BIOSYNTHESIS: A MECHANISTIC OVERVIEW

Cholesterol is an essential component of cell membranes and the precursor for the synthesis of steroid hormones and bile acids. The synthesis of this molecule occurs partially in a membranous world (especially the last steps), where the enzymes, substrates and products involved tend to be extremely hydrophobic. The importance of cholesterol increased in the last half century due to its association with cardiovascular diseases, which are considered one of the top leading causes of death worldwide. Facing the current need for new drugs capable of controlling the levels of cholesterol in the bloodstream, it is important to understand how cholesterol is synthesized in the organism and identify the main enzymes involved in this process. Taking this into account, this review presents a detailed description of several enzymes involved in the biosynthesis of cholesterol. In this regard, the structure and the catalytic mechanism of the enzymes involved in the cholesterol biosynthesis, from the initial 2-carbon acetyl-CoA building block, will be reviewed and their current pharmacological importance discussed. We believe that this manuscript may contribute to a finer level of understanding of cholesterol metabolism and that it will serve as a useful resource for future studies of the cholesterol biosynthesis pathway.

Adapted from reference [181]

In this review, Diana Gesto wrote several parts including sections 4.2.3 and 4.2.10, reviewed the whole manuscript and did some of the figures.

4.1. Introduction

Cholesterol is the major sterol present in animal tissues. This molecule is almost planar and rigid and contains a steroid nucleus of four fused rings, three of which with six carbons and a forth with five. The molecule is amphiphilic, having a hydrophobic hydrocarbon body and a hydrophilic hydroxyl head group.

In mammals, cholesterol plays a vital role in life, being an essential component for the normal functioning of cells. Its roles range from component in cell membranes to precursor of several steroid hormones.

Cell membranes are very complex systems, which separate the cells cytosol from the external medium or from the medium inside cell compartments (like lysosome or peroxisomes), while still allowing for the transfer of compounds from the inside to the outside and vice-versa, as well as carrying other important functions. Still, in a simplistic manner, they can be seen essentially as a double layer of phospholipids. Cholesterol is normally present in membranes and its importance can be easily understood just by comparing the difference in fluidity of membranes with different cholesterol concentrations. The polar head group of cholesterol interacts with the similar head group of phospholipids, while the nonpolar group interacts with their hydrophobic tails. Since cholesterol is somewhat rigid (more rigid than phospholipids), membranes with larger cholesterol content will tend to be more rigid and packed, while those with less cholesterol will be more fluid. Cholesterol is also important in other membrane processes, such as endocytosis.

In addition to being a structural component of membranes, cholesterol also serves as a precursor for the biosynthesis of several compounds like bile acids, vitamin D and several steroid hormones that are produced by the adrenal gland and by the male and female sex glands.

Over the years cholesterol has gained a bad reputation in the world of health and nutrition, especially because of its association with cardiovascular diseases. According to the World Health Organization (WHO), from the top 10 leading causes of death worldwide in 2008, ischemic heart diseases was number one, accounting for 12.8% of deaths, followed by stroke and other cerebrovascular disease as number two (10.8%). Since both of these diseases are associated with high levels of cholesterol in the blood (hypercholesterolemia), it is easy to understand why this molecule has acquired such a bad name.

Despite its association with several heart conditions, cholesterol should not be perceived as a bad compound that needs to be avoided at all costs. In fact, life as we know it would not be possible without cholesterol.

In terms of composition, it is possible to distinguish between good and bad cholesterol. Since it is a lipid, cholesterol cannot be dissolved in the bloodstream, which is water-based. To get around this problem, the body packages cholesterol and other fats into protein-covered molecular assemblies called lipoproteins that do mix easily with

blood. The proteins that are used in this assembly are known as apolipoproteins. When the proportion of protein to lipids (cholesterol and others) in the lipoprotein is high, they are known as high-density lipoprotein (HDL), or good cholesterol. On the other hand, when this proportion is small, meaning that there is a larger lipidic content, they are called low-density lipoprotein (LDL), or bad cholesterol.

The high levels of cholesterol are especially dangerous when the concentration of LDL in the blood is high and that of HDL is low. High levels of LDL may lead to atherosclerosis, which is one of the main causes for both ischemic heart disease and cerebrovascular diseases[182-183]. Atherosclerosis is nothing more than the accumulation of fatty materials, such as cholesterol, in the blood vessels. It is a complex process, which involves a chronic inflammatory response to oxidized LDL on the walls of arteries. LDL is very rich in cholesterol and cholesteryl esters, which, when oxidized, are toxic to the cells on the walls of arteries, triggering an inflammatory response. This leads to a pathogenic accumulation of cholesterol in blood vessels and the formation of atherosclerotic plaques, resulting in the constriction of blood vessels. Atherosclerosis occurs when the amount of cholesterol in the blood, due to either unregulated synthesis or considerable ingestion of cholesterol rich foods, exceed the amount needed for the production of steroids, bile acids and membranes[8].

Hypercholesterolemia is currently treated with a combination of dietary and pharmaceutical therapies[184-185]. Often, more than a single pharmaceutical agent and a dietary regimen are necessary to decrease total cholesterol and LDL levels to the desired level. Drugs such as bile acid sequestrates, niacin and statins are commonly used to treat hypercholesterolemia and atherosclerosis. The application of several of these compounds is, however, limited by the numerous side effects that can be experienced by patients. Thus, the need for therapeutic agents that would decrease cholesterol levels still exists nowadays.

Facing the current need for new drugs capable of controlling the levels of cholesterol in the bloodstream, it is important to understand how the organism synthesizes this molecule and identify the main enzymes involved in this process. Taking this into account, the most important enzymes participating in the cholesterol biosynthesis will be described in this review and special attention will be given to those that are current drug targets.

4.2. Enzymes involved in the Cholesterol Pathway

All cholesterol present in our bodies arises from two different sources: it can be either synthesized de novo within our cells, or obtained through ingestion of certain foods. Although many people regularly include these foods in their diet, there is no need to ingest them for the sole purpose of obtaining cholesterol, since our own cells are capable of producing enough quantities of this molecule for our bodily requirements[186]. Nonetheless, whether or not there is dietary intake of cholesterol, its levels are maintained through regulation of the synthesis and absorption, which means that when low quantities of cholesterol are ingested, absorption and synthesis will be upregulated. Likewise, if the dietary intake is high, its excretion will be increased and its rate of synthesis will be decreased.

The biosynthesis of cholesterol is a complex process, heavily regulated at several points throughout its progression. Some of the intermediaries can be diverted and used as precursors in the biosynthesis of other compounds or perform themselves certain functions on the body. This process requires numerous enzymes, some of which are accounted amongst the most regulated enzymes currently known.

The first step in the synthesis of cholesterol is the formation of mevalonate from acetate. It begins with the condensation of two acetyl coenzyme A (acetyl-CoA) molecules to form acetoacetyl-CoA, a process catalysed by the enzyme thiolase. Next, HMG-CoA synthase catalyses the reaction between acetoacetyl-CoA and another molecule of acetyl-CoA in order to form HMG-CoA. The final step in the synthesis of mevalonate is accomplished by HMG-CoA reductase. This step is not only the committed step of the whole process, but also the rate-limiting one.

The subsequent step in the biosynthesis of cholesterol comprises the conversion of mevalonate into two activated isoprenes (isopentanyl-5-pyrophosphate and dymethylallyl pyrophosphate). Following a series of successive condensations of activated isoprenes, a 30 carbon molecule, squalene, is formed. Squalene is the biochemical precursor of all steroids, and despite being a linear compound, its structure can still be linked to that of cyclic steroids. In order to form cholesterol, squalene has to endure a succession of changes, being initially converted to lanosterol, a four-ring compound, which is finally transformed into cholesterol after several sequential reactions.

Figure 20 describes the overall process by which cholesterol is synthesized, from acetyl-CoA to lanosterol. Each of these reactions will be described in more detail in the next sections.

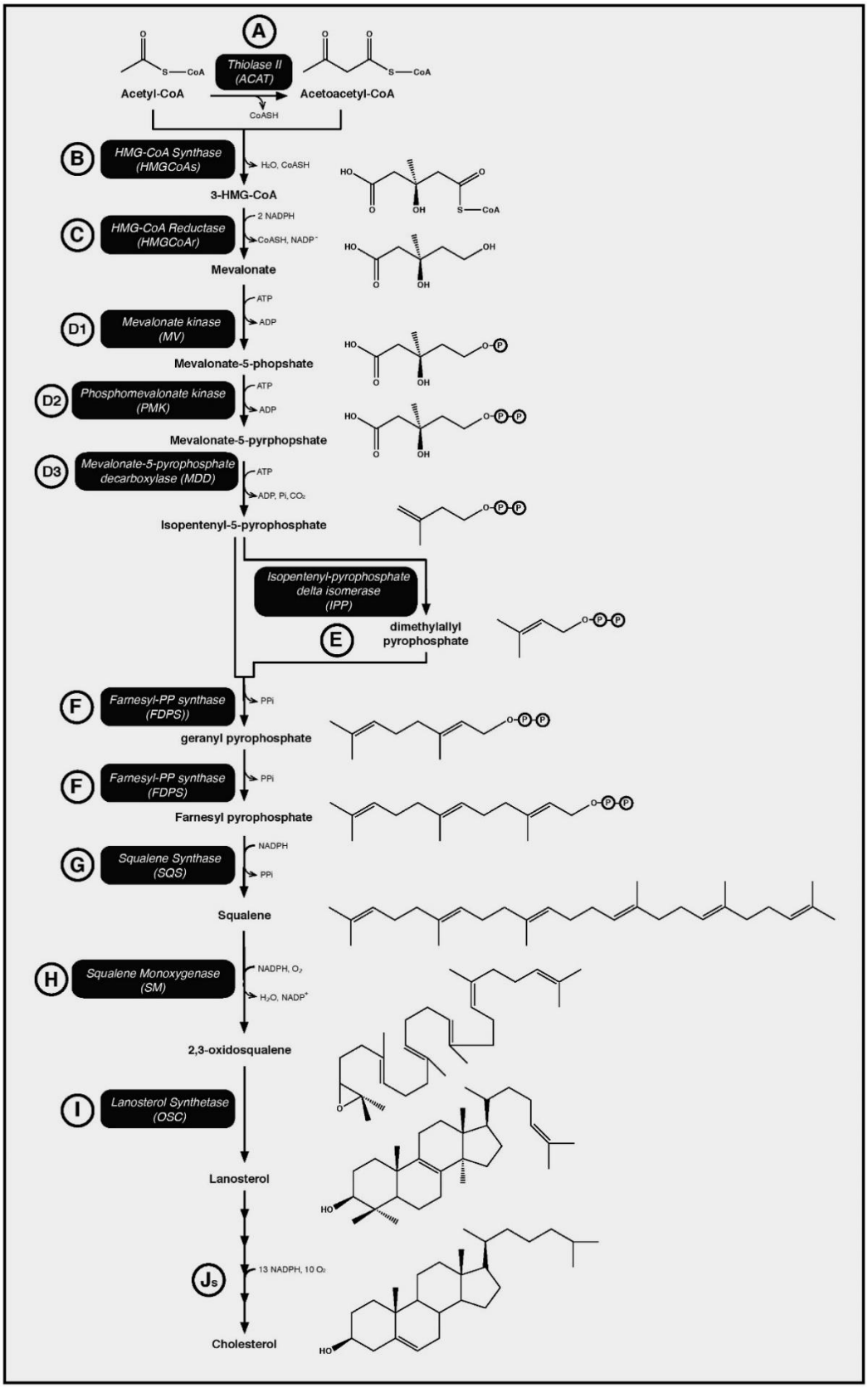


Figure 20 - Diagram describing most of the enzymes involved in the cholesterol biosynthesis. Each enzyme is identified with a different letter that corresponds to the header of the following sections, where each enzyme is describe in more detail. When this letter is followed by an "s" it means that multi enzymes are involved in that step.

4.2.1. Thiolase

Thiolases, also known as acetyl-coenzyme A acetyltransferases (ACAT), are the enzymes responsible for the conversion of two units of acetyl-CoA in acetoacetyl-CoA. The acetyl-CoA is either obtained from an oxidation reaction (e.g., fatty acids or pyruvate) in the mitochondria or synthesized from the cytosolic acetate derived from cytoplasmic oxidation of ethanol, which is initiated by cytoplasmic alcohol dehydrogenase.

Thiolases are ubiquitous enzymes that have key roles in many vital biochemical pathways[187]. The members of this family of enzymes can be divided into two broad categories: degradative thiolases (3-ketoacyl-CoA thiolase, (EC 2.3.1.16)) and biosynthetic thiolases (acetoacetyl-CoA thiolase, (EC 2.3.1.9)). 3-ketoacyl-CoA thiolases have broad chain-length specificity for its substrates and are involved in degradative pathways such as fatty acid beta-oxidation[188]. Acetoacetyl-CoA thiolases are specific for the thiolysis of acetoacetyl-CoA and are involved in the first step of many biosynthetic pathways, including those that generate cholesterol, steroid hormones and ketone body energy storage molecules[189]. There are two types of acetoacetyl-CoA thiolases, the mitochondrial ACAT1 and the cytosolic ACAT2. Kovacs et al. suggest a possible distribution of ACAT1 between peroxisomes and mitochondria, as experimental evidence supports the formation of acetoacetyl-CoA in peroxisomes[190].

The physiological importance of thiolases can be illustrated by the severe, and usually lethal, phenotypes of patients with deficient thiolases[191-192], particularly in the case of defective biosynthetic thiolases[193-194]. Currently, inhibitors against ACATs have been used to develop new drugs against African sleeping sickness and other diseases[195-196].

Most enzymes of the thiolase family are dimers. The active site of the biosynthetic thiolases is located in a shallow cleft at the protein surface. Enzyme kinetics, active-site labeling and site-directed mutagenesis experiments allowed the identification of three important residues, Cys89, His348 and Cys378 , which have been found to be important for the catalytic activity of the enzyme (Figure 21-A). These residues are part of an extensive hydrogen bond network, which stretches from the active site to the enzyme backside, suggesting that this hydrogen bond network is important for the stabilization of different protonation states of the catalytic residues[188].

On the basis of these studies, a ‘ping-pong’ reaction mechanism consisting of two steps has been proposed[197] (Figure 21-B). The first step of the mechanism involves the deprotonation of Cys89 by His348 (Figure 21-B step 2)[198]. Simultaneously, the nucleophilic Cys89 attacks the acyl-CoA substrate, leading to the formation of a covalent

acyl-CoA tetrahedral intermediate (Figure 21-B step 2). It has then been proposed that afterwards Cys378 becomes negatively charged, following the donation of its proton to the leaving group of the acetylation reaction, CoA (Figure 21-B step 3). Subsequently, in the second half of the synthesis reaction, Cys378 acts as a catalytic base which deprotonates carbon C2 of a new acetyl-CoA molecule, resulting in the formation of an enolate intermediate (Figure 21-B, step 4). The following steps of the catalytic mechanisms are the reverse of step 2 and 3 and involve the formation of a second covalent tetrahedral intermediate (Figure 21-B step 5), the transference of a proton from His348 to Cys89 and the subsequent release of the product of the reaction (Figure 21-B step 6 and 7). During the catalytic process, the formation of the negatively charged tetrahedral intermediates are stabilized the oxyanion hole, which is composed by the Asn316-Wat82 dyad and His348.

4.2.2. HMG-CoA Synthase

HMG-CoA synthase (HMG-CoA-S, E.C. 2.3.3.10) is a 42 kDa homodimeric protein that catalyses the irreversible condensation of acetyl-CoA and acetoacetyl-CoA to form 3-hydroxy-3-methylglutaryl–CoA (HMG-CoA) and CoA.

HMG-CoA-S is found in eukaryotes, archaea and certain bacteria[199]. Cells of higher eukaryotes have three forms of HMG-CoA synthase: the cytosolic, the mitochondrial and the peroxisomal[200-202]. Cytosolic and peroxisomal HMG-CoA-S (cHMG-CoA-S and pHMG-CoA-S) catalyse the second step in the synthesis of cholesterol. This starts the isoprenoid pathway, from which results cholesterol and other important products, such as ubiquinone, dolichol, isopentenyl adenosine and farnesyl groups. Mitochondrial HMG-CoA-S (mHMG-CoA-S) participates mainly in the ketogenesis pathway for acetoacetate biosynthesis. Acetoacetate is then transformed into hydroxybutyrate and acetone, often known as ketone bodies. In bacteria, isoprenoid precursors are generally synthesized via an alternative non-mevalonate pathway. However, a number of Gram-positive pathogens use a mevalonate pathway involving a bacterial HMG-CoA-S isoform that is parallel to that found in eukaryotes[203].

The main structure of HMG-CoA-S can be divided in two main regions: the larger upper region that contains a conserved αβαβα structure from the thiolase fold, and the smaller lower region, which is unique among HMG-CoA-S enzymes and consists of several α-helices and β-strands[204]. The interface between the upper and lower regions defines a 15 Å-deep narrow tunnel that forms the active site in each monomer (Figure 22-A).

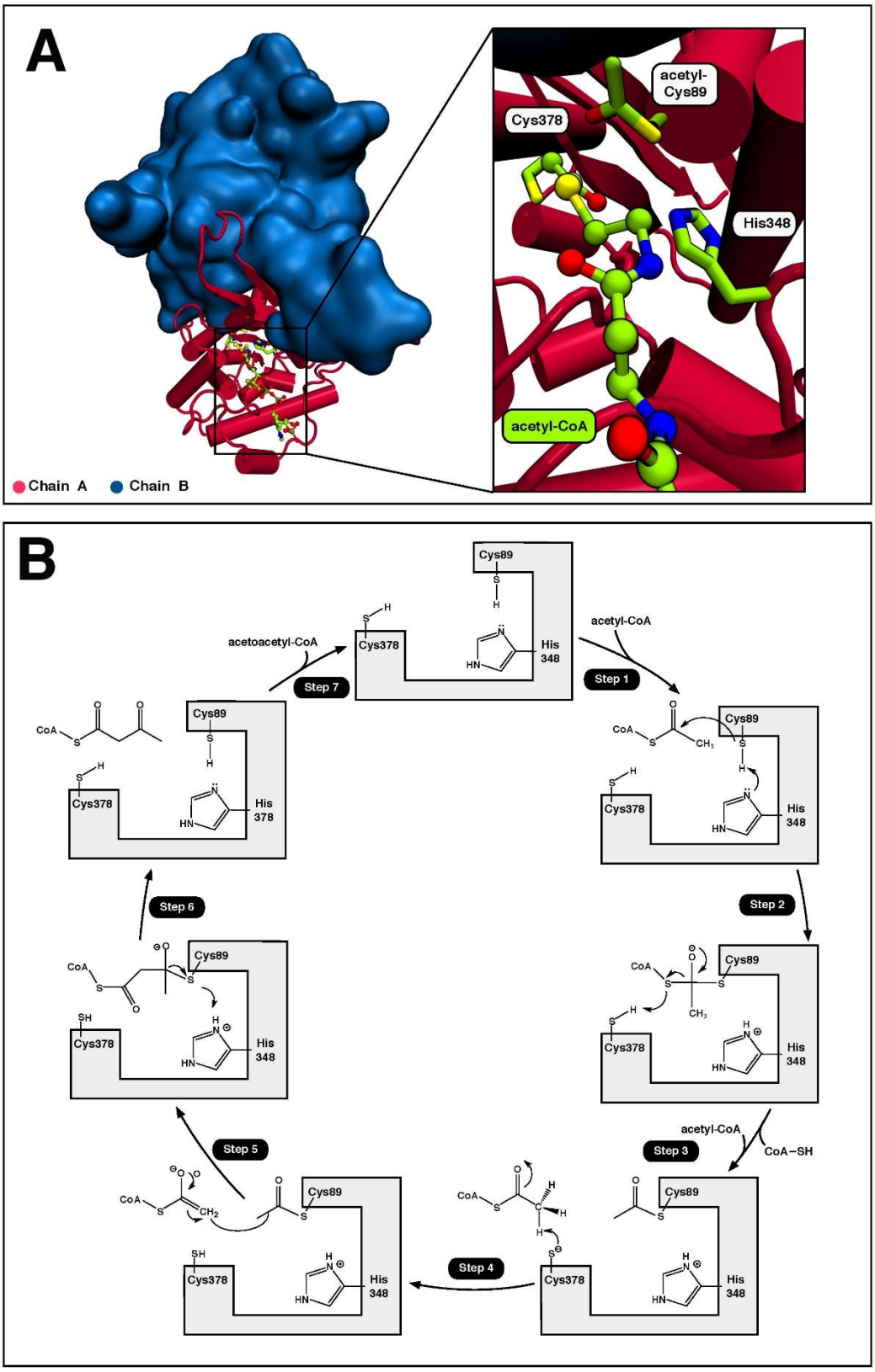


Figure 21 - A: Structure of Thiolase II from Zoogloea ramigera (pdb code 1DM3[188]) and the active site with a reaction intermediate (the enzyme is acetylated at Cys89 and a molecule of acetyl-CoA is found in the active site pocket). B: Proposed catalytic mechanism[197].

Both substrates, acetyl-CoA and acetoacetyl-CoA, fit in the active site tunnel with the nucleotide part pointing toward the protein surface. The pantetheine part of the substrate is located in the centre of the tunnel that is enclosed by a hydrophobic sleeve, and the CoA thiol group is buried in the bottom of the active site tunnel, where the catalytic triad, formed by Glu95, Cys129 and His264, is located[204](Figure 22-A). Cys129 is placed in the deepest portion of the active site and is responsible for the nucleophilic attack on the substrate. The main function of His264 and Glu95 is to remove the proton from the Cys129 and turn it into a stronger nucleophile. Mutations on any of these residues inactivate the enzyme[205-207].

From a mechanistic point of view, the condensation reaction catalyzed by HMG-CoA-S is similar to that of thiolase. However, the mechanism diverges on the second (condensation) step of the reaction, where the methyl group of the acetylated enzyme is activated and attacks the incoming β-keto thioester, whereas the members of the thiolase family activate a carbon of the second substrate to attack the enzyme-bound thioester[208]. Based on these mutagenic studies, other kinetic studies[207, 209] and co-crystallized X-ray structures with bound intermediates[204, 208, 210], one can propose the mechanism of HMG-CoA synthase which is summarized in Figure 22-B.

The first step of the catalytic process involves the activation of Cys129, through the action of His264, which abstracts a proton from that residue (Figure 22-B, step 2). The negatively charged cysteine promptly attacks the acetyl-CoA and forms a thioester acyl−enzyme intermediate at the same time that a reduced CoA is generated. The next step involves a Claisen-like condensation of the second substrate of the reaction, acetoacetyl-CoA, with the thioester acyl−enzyme intermediate to form HMG-CoA, while it still retains the thioester bond to the enzyme (Figure 22-B, step 3). This reaction is favored by the close presence of Glu95, which abstracts a proton from the methyl group of the thioester acyl−enzyme intermediate, turning it into a better nucleophile, and by His264, which gives a proton to one of the carbonyl groups of acetoacetyl-CoA, turning it into a better electrophile [204, 211]. The last two steps of the catalytic process involve the hydrolysis of the acetylcysteine bond (Figure 22-B step 4 and 5), from which results the free HMG-CoA. In the end of this step Cys129 becomes reduced and it is suggested that Glu95 loses a proton to the solvent, thus making the active site ready for a new catalytic cycle[211].

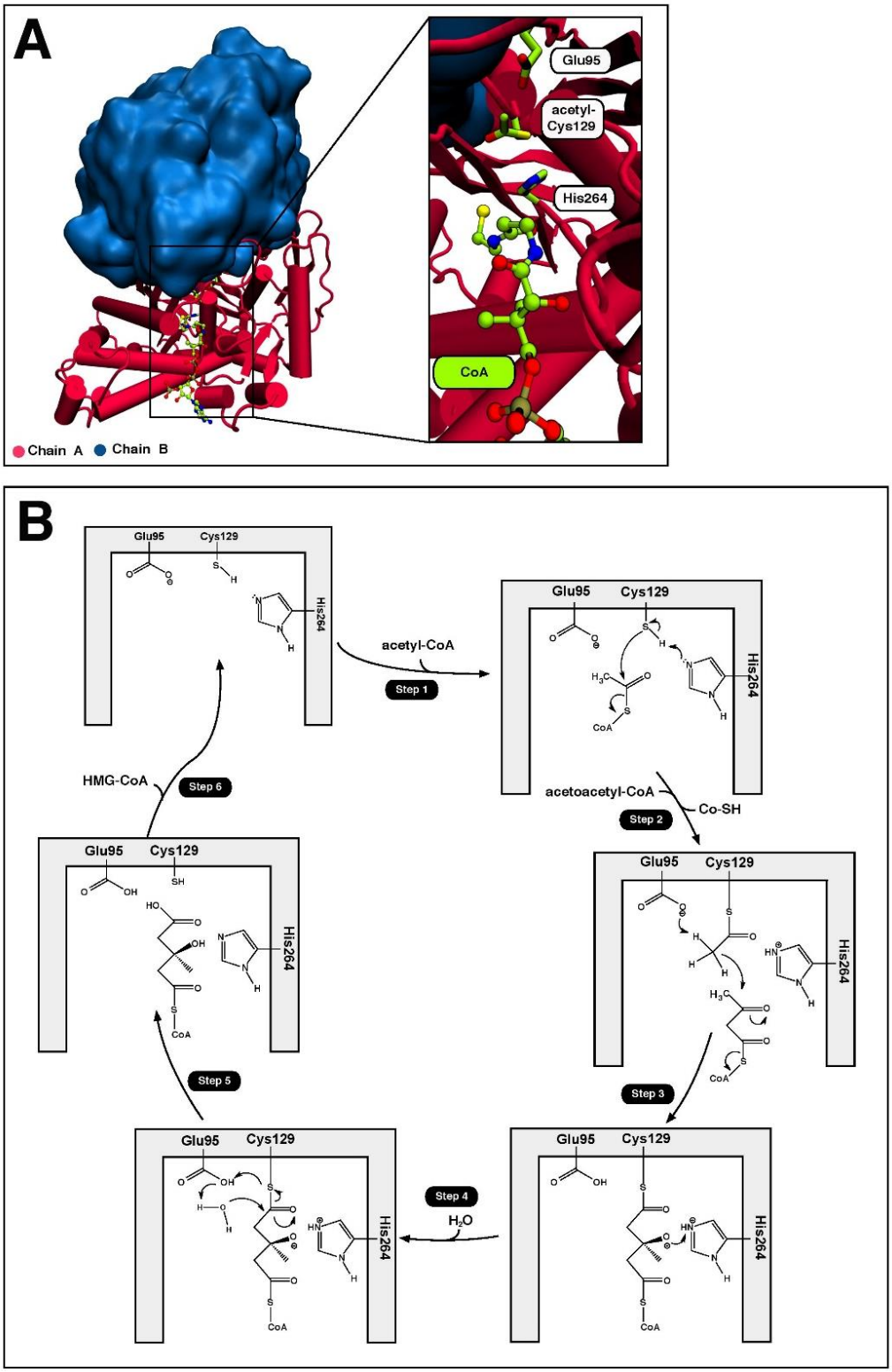


Figure 22 - A: Structure of cHMG-CoA synthase (PDB code 2P8U[204]) and the active site with the product of the reaction and Cys117 acetylated. B: Proposed catalytic mechanism of the enzyme[204 205-207].

HMG-CoA-S is currently a potential drug target. Inhibitors addressed to this enzyme can be used to regulate the serum cholesterol level or even in the development

of novel antibiotics against certain pathogens. This last area of research goes in line with the discovery that many human pathogenic gram-positive bacteria can only produce isopentoyl disphosphate (IPP) through the mevalonate pathway, which is a necessary pathway for bacteria to live. Taking into account the differences between the prokaryotic and eukaryotic HMG-CoA-S, this suggests that the inhibition of the mevalonate pathway in bacteria may not affect the HMG-CoA-S enzymes of the host body. This could thus lead to a new generation of antibacterial drugs that have a significant therapeutic advantage over the commonly used broad-spectrum antibiotics[212-213].

mHMG-CoA-S has also an important role in diabetes. The reaction catalyzed by this enzyme is over-activated in patients with diabetes mellitus type I and if left untreated, due to prolonged insulin deficiency and the exhaustion of substrates for gluconeogenesis, results in the shunting of excess acetyl-CoA into the ketone synthesis pathway and the development of diabetic ketoacidosis. This means that under such prognostic, inhibitors of this enzyme could help in the reversion of the process and decrease in the levels of ketonic compounds in blood serum.

4.2.3. HMG-CoA Reductase

3-hydroxy-3-methyl-glutaryl-CoA reductase or HMG-CoA reductase (HMG-CoA-R) catalyses the NADP-dependent (in mammals) or NAD-dependent (in prokaryotes) synthesis of mevalonate from HMG-CoA[190, 214]. For this reason the enzyme commission designated this enzyme as EC 1.1.1.34 for the NADPH-dependent enzyme, whereas 1.1.1.88 links to the NADH-dependent enzyme.

From sequence analysis it was possible to divide different HMG-CoA-R in two main classes. Class I enzymes comprise those found in eukaryotes and the majority of archaea, while class II enzymes are found in prokaryotes and some archaea.

Class I HMG-CoA-R contains a transmembrane domain (lacking in archaeal enzymes), and a C-terminal catalytic region[215]. The catalytic region of these enzymes is well conserved, with sequence identities of ~60%, and can be subdivided into three domains: an N-domain (N-terminal), a large L-domain, and a small S-domain (inserted within the L-domain). The L-domain binds the substrate, while the S-domain binds NADPH. The membrane region is diverse in sequence and length, and exhibits functional differences. For instance plants have two membrane domains, yeast seven and mammals eight[216]. Class II HMG-CoA-R lacks the membrane domain and so it is found dissolved in the cytosol. Their catalytic region is structurally related to the one found in

class I, but it consists of only two domains: a large L-domain and a small S-domain (inserted within the L-domain). As with class I enzymes, the L-domain binds the substrate, whereas the S-domain binds NADH (instead of NADPH in class I).

HMG-CoA-R from both classes has a dimeric active site (Figure 23–A), with residues contributed by both monomer, and a nucleotide-binding motif that is found in many enzymes that use the dinucleotides NADH or NADPH for catalysis[217]. The core regions containing the catalytic domains of the two enzymes have similar folds. Despite the differences in amino-acid sequence and overall architecture, functionally similar residues participate in the binding of coenzyme A, and the position and orientation of four key catalytic residues (glutamate, lysine, aspartate and histidine) is conserved in both classes of HMG-CoA-R.

Both classes of HMG-CoA-R are catalytically competent and have similar reaction mechanisms and kinetic parameters[216, 218-220]. The currently available proposal for the catalytic mechanism of HMG-CoA-R is based on kinetic and labeling experiments, site-directed mutagenesis, protein sequencing and in the mechanism of dehydrogenases, which catalyze a similar reaction[221-225226].

The overall reaction catalyzed by these enzymes can be divided into three main stages (Figure 23-B). Once the HMG-CoA and NADH (or NADPH) bind to the active site, the thioester of HMG-CoA-R is reduced and the generated intermediate is stabilized by the protonated Glu559 and Lys691 (Figure 23-B step 1 and 2). By the end of this reaction, a mevaldyl-CoA hemi-thioacetal intermediate is generated, and the NAD^+ (or $NADP^+$) cofactor dissociates from the active site. In the second stage of the mechanism, Glu559 assists the base-catalyzed decomposition of the hemithioacetal to produce mevaldehyde and the CoA thiolate anion, which is promptly protonated by the cationic His752 (Figure 4-B step 3). In the third stage of the catalytic mechanism, the mevaldehyde is reduced by a second NADH (or NADPH) molecule and the intermediate is again stabilized by the protonated Glu559 and Lys691 (Figure 23-B Step 3). Subsequently, the product mevalonate, CoASH, and the oxidized cofactor (NAD^+ or $NADP^+$) are released and the enzyme is ready for a new turnover (Figure 23-B Step 5). During the full catalytic process, Asp690 established an important hydrogen bond with Lys691 that has been shown to be important for the catalytic process[227].

The reaction catalyzed by HMG-CoA-R is the point of feedback control for the mevalonate pathway, which is responsible for the biosynthesis of many important isoprenoid compounds in humans. This enzyme is also one of the most regulated enzymes in our body[10]. Dietary cholesterol exerts feedback control mostly at a

translational level[228-230], but regulation mechanisms at the transcription level[20-21], enzyme degradation[23, 231] and modulation of enzyme activity[24] are also known.

Because of its importance, this enzyme is also a main target for medical intervention. Currently, this enzyme is the target of a class of drugs called statins, which are used to treat hypercholesterolemia and reduce the risk of cardiovascular disease. Additionally, this enzyme has been a target for the development of new antimicrobial agents, which take advantage of the evolutionarily divergence between the HMG-CoA-R of pathogenic bacteria (from Class II) and that of eukaryotes (Class I).

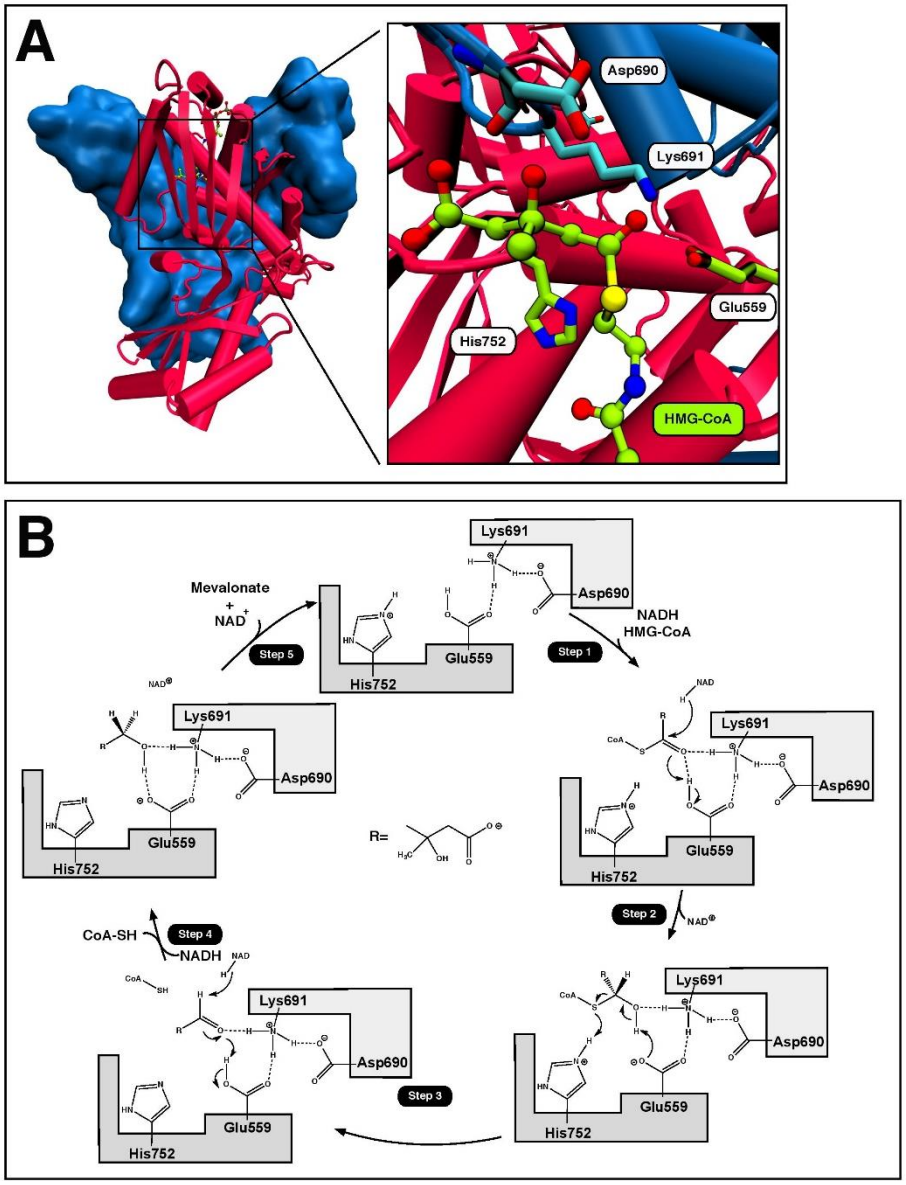


Figure 23 - A: Structure and active site of HMG-CoA-R (pdb code 1DQ9[216]). B: Currently accepted catalytic mechanism of HMG-CoA-R[221-225, 232].

Statins are potent competitive inhibitors of HMG-CoA-R[37] and can be divided in two different types based on their structure[38]. Type I statins (lovastatin, pravastatin and simvastatin) are natural fungal products, whereas type II statin are fully synthetic, characterized by the presence of larger hydrophobic regions and attached fluoro-phenyl groups[40]. The ability of statins to inhibit HMG-CoA-R arises from their HMG-like moiety, which competes with HMG-CoA to bind to the HMG binding site of the enzyme (see pdb codes 1HW8[38] and 1HMK[38] where HMG-Coa-R is bounded to mevastatin and atorvastatin, respectively). Although the hydrophobic part of statins is normally very different from that of the coenzyme A portion of the substrate, these non-polar groups also contribute to block the access of HMG-CoA to the active site. The affinity of the enzyme for statins is slightly higher than its affinity for the substrate[41].

HMG-CoA-R can also be inhibited at a translational level. The oxylanosterol 15-oxa-32-vinyl lanost-8-ene-3b,32 diol is one of the inhibitors that work at this level and it is a potent hypocholesterolemic agent[233].

New antimicrobial agents that target HMG-CoA-R of pathogenic bacteria (class II) are also a hot topic of research in this field, in part due to the increasingly problematic broad antimicrobial resistance among gram-negative bacteria to the current available drugs[234]. Very few novel pharmacologic agents are in the research and development pipeline and several recent studies have shown that some statins exhibit antibacterial action[235-236].

4.2.4. ATP-dependent Enzymes involved in the Cholesterol Pathway.

In animal cells, the cholesterol biosynthetic pathway contains a unique series of three sequential ATP-dependent enzymes that convert mevalonate to isopentenyl diphosphate: mevalonate kinase (MK), phosphomevalonate kinase (PMK), and mevalonate 5-diphosphate decarboxylase (MDD). According to Hogenboom and co-authors, the enzymes MK, PMK and MDD are cytosolic enzymes[237-239], but this contradicts recent studies of Kovacs and co-authors, which confirm their previous findings of peroxisomal localization of the three enzymes using stable isotopic techniques and human cells[190, 240-241].

4.2.4.1. Mevalonate Kinase

Mevalonate kinases (EC 2.7.1.36) catalyze the transfer of the γ-phosphoryl group of ATP to the C5 hydroxyl oxygen of mevalonic acid to form mevalonate 5-phosphate, a key intermediate in the biosynthetic pathway for isoprenoids and sterols from acetate[242-243].

The enzyme MK was discovered in the late 1950s in yeast[244], but it suffered from more than three decades of neglect, as research on the isoprenoid pathway was mainly focused on HMG-CoA-R. The interest in MK has been revived, however, because this enzyme was found to be involved in the synthesis of diverse non-sterol isoprenoid metabolites that participate in numerous cellular functions, e.g., protein prenylation, protein glycosylation, and cell cycle regulation. This enzyme is also currently being studied as a promising anticancer target since certain types of cancer are particularly sensitive to mevalonate pathway inhibition[245]. In addition, the significance of MK has been further highlighted by the implication of the enzyme in human inherited diseases, such as mevalonic aciduria and hyperimmunoglobulinemia D/periodic fever syndrome.

MK can be found in eukaryotes, archaebacteria, and some eubacteria. The mammalian enzyme is reported to be a homodimer with a subunit molecular mass of 42 kDa[246].

Kinetic studies suggest that the enzyme catalyzes an ordered sequential reaction, with mevalonic acid binding prior to the enzyme in relation to ATP and the phosphomevalonate being the first product of the reaction[247]. Site-specific mutagenesis studies have been employed to identify catalytic residues and to investigate the mechanism of the enzyme. Asp204 has been suggested as the catalytic base that abstracts the proton from the C5-OH group of mevalonic acid[246]. Ser146 and Asp204, located in a conserved glycine-rich region, have been implicated in the binding of Mg-ATP[248]. Lys13 has been shown to be involved in the binding of ATP as well as facilitating catalysis[249](Figure 24-A).

The reaction mechanism has not yet been deeply studied. However, the reaction is likely to be initiated by the nucleophilic attack of the 5′-OH of mevalonate to the γ-phosphorus of ATP. At the same time, one proton is transferred from mevalonate to Lys13 and another from Lys13 to the phosphate group (Figure 24-B-step 1). Based on mutation data available for human MK, it is likely that Lys13 or even Asp204 may function as a general base in this process, and abstract a proton from the 5′-OH of mevalonate, thereby facilitating the nucleophilic attack of mevalonate[250]. The Mg^{2+} is coordinated by

Glu193 and Asp204 (being this one stabilized by the close Ser146), which are suggested to be responsible for the activation of the γ-phosphate.

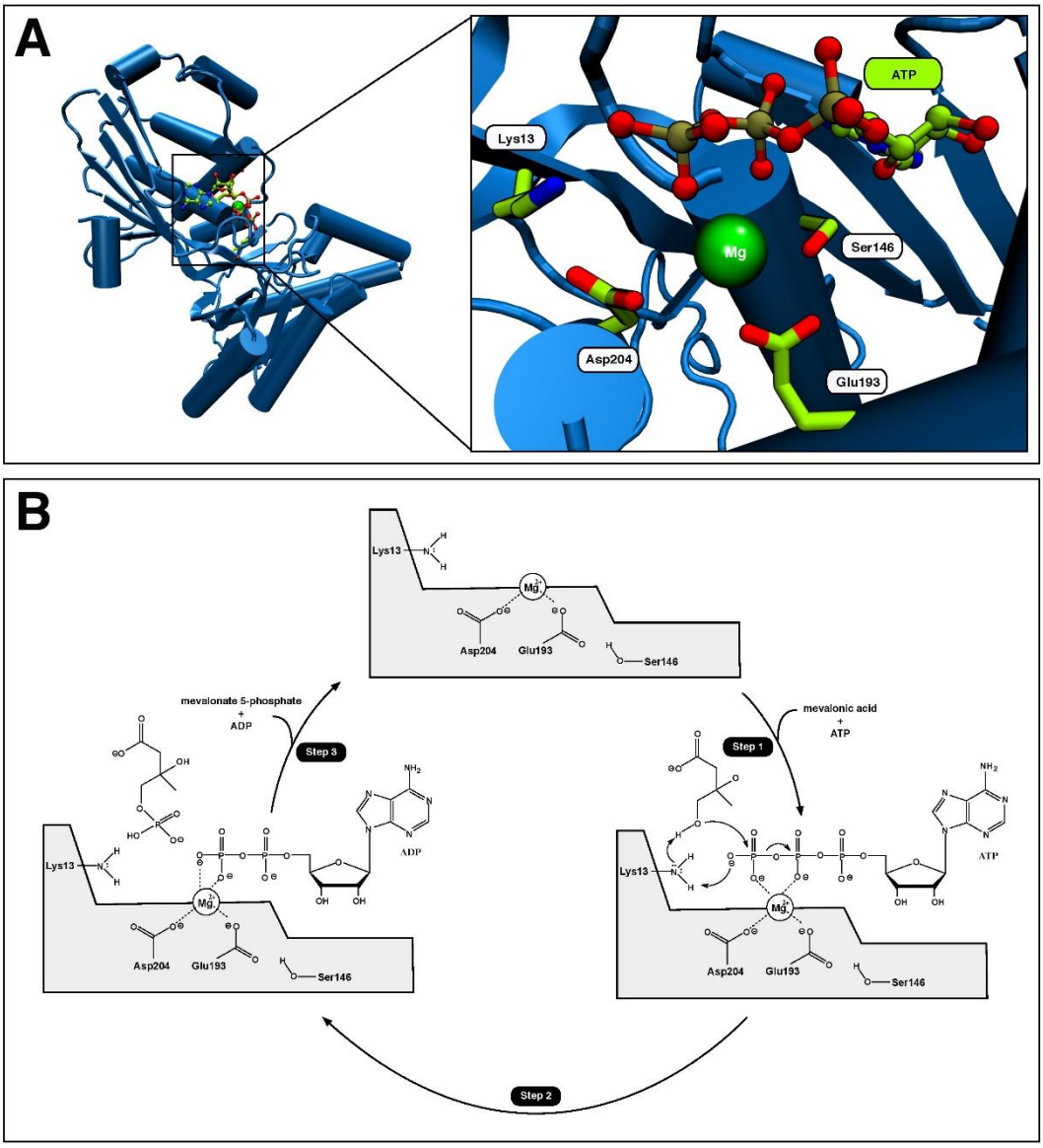


Figure 24 - A: Structure and active site topology of mevalonate Kinase (pdb entr 1KVK[242]). B: Schematic proposal for the catalytic mechanism of mevalonate kinase[250].

Currently, most of the inhibitors targeting MK have in their structure phosphate groups. For example, farnesyl diphosphate and related compounds are potent and competitive inhibitors of MK. They bind in the active site similarly to what is observed with ATP, through the same type of interactions that were described above (see PDB code: 2R42[243]).

4.2.4.2. Phosphomevalonate Kinase

Phosphomevalonate kinase (EC 2.7.4.2) catalyzes the transfer of a second γ-phosphoryl group from ATP to mevalonate 5-phosphate (MVAPP), which results in the formation of ADP and mevalonate 5-diphosphate.

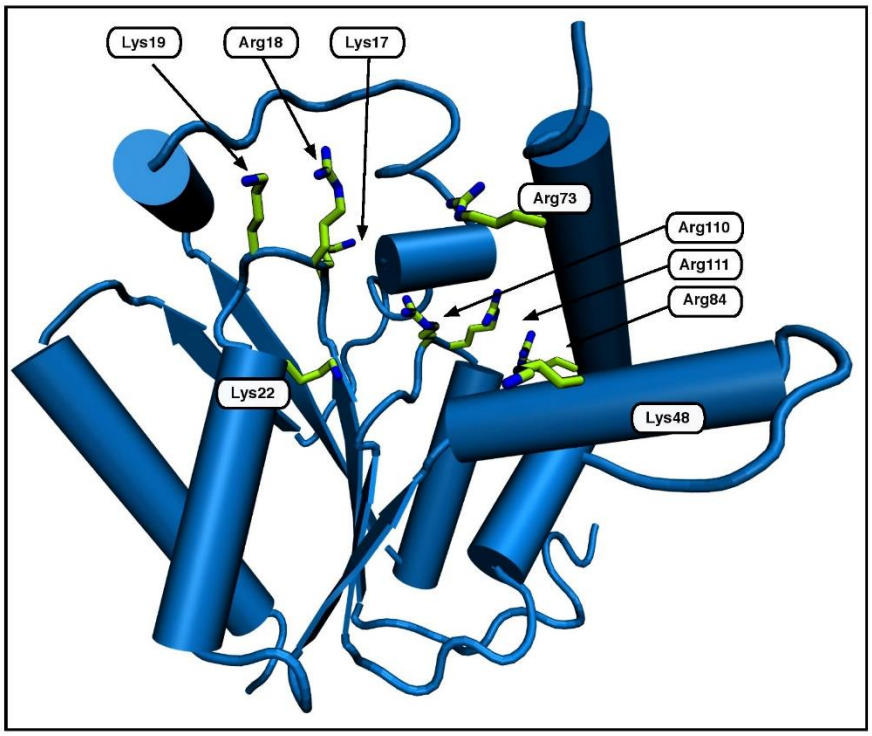


Figure 25 - Structure of phosphomevalonate kinase (PMK) and some important active site residues that have been identified by experimental means (PDB code: 3CH4[251]).

The crystal structure of the human phosphomevalonate kinase has 192 amino acid residues and a predicted molecular weight of 21.8 kDa. Despite its importance on the biosynthetic source of a diverse class of metabolites, the mechanism of PMK is not completely characterized. The only PDB structure of PMK that is currently available on the protein databank (Figure 25, PDB code: 3CH4[251]) reveals a positively charged cavity on the protein surface that is believed to be essential for the binding of the phosphate group and hydroxyl groups of the substrate. This cavity includes the residues Lys48, Arg73, Arg84, Arg110, Arg111 and Lys73[252-253]. Site-directed mutagenesis studies have also identified four additional amino acid residues, Lys17, Arg18, Lys19, and Lys22 that, if mutated, decrease the catalysis up to 10,000-fold and are believed to be part of the ATP binding motif[254-256]. This means that these residues are directly involved in the

catalysis and therefore in the binding and transfer of the γ-phosphoryl group of ATP to mevalonate 5-phosphate.

Presently, no co-crystallized structure of PMK with the substrate, reaction intermediate or inhibitor is available. Thus, no absolute conclusions about the catalytic mechanism can be inferred. The current understanding of the mechanism considers that bringing two phosphate groups in proximity to react is especially challenging, given the high negative charge density on the four phosphate groups in the active site. Therefore, only a very specific and highly positively charged active site, like the one found in PMK, can catalyze this sort of reaction. Furthermore, the possible involvement of one or more magnesium ions should not be discarded, as they could help the binding of the phosphate groups as well as the catalysis, alike to what is observed in the previously described kinases.

4.2.4.3. Diphosphomevalonate Decarboxylase

Diphosphomevalonate decarboxylase, also known as mevalonate diphosphate decarboxylase (MDD, EC 4.1.1.33), catalyses the final step of the mevalonate pathway, i.e., the divalent cation-dependent decarboxylation of MVAPP to isopentenyl diphosphate (or isopentenyl pyrophosphate, IPP), with concurrent hydrolysis of ATP to form ADP and inorganic phosphate. This reaction is required for the production of polyisoprenoids and sterols from acetyl-CoA.

Inhibition of this enzyme effectively diminishes biosynthesis of cholesterol[257] 60 and low MDD activity correlates with decreased cholesterol levels in a hypertensive rat strain[258]. Genes encoding this enzyme have been detected in archaebacteria, some eubacteria, protozoa, plants, fungi and animals.

The X-ray structure of human MDD shows that it consists of two domains (Figure 26-A) and has many similarities to the structures of bacterial, protozoan and yeast[259-260]. This structure revealed the presence of tightly bound PO_4^- and SO_4^- anions, suggesting that they might be important for the catalytic process. Site-directed mutagenesis studies also revealed four conserved residues that are important for catalysis, Asp17, Ser127, Arg161 and Asp305[259, 261-262]. Asp17 and Arg161 are believed to be part of the phosphoryl acceptor-binding site. Asp305 and Ser127 are suggested to be actively involved in the base catalysis that takes place in the active site. The correct position of the magnesium ion in the active site is still not fully established. Based on these studies,

the catalytic mechanism of MDD has been proposed to be as it is illustrated on Figure 26-B.

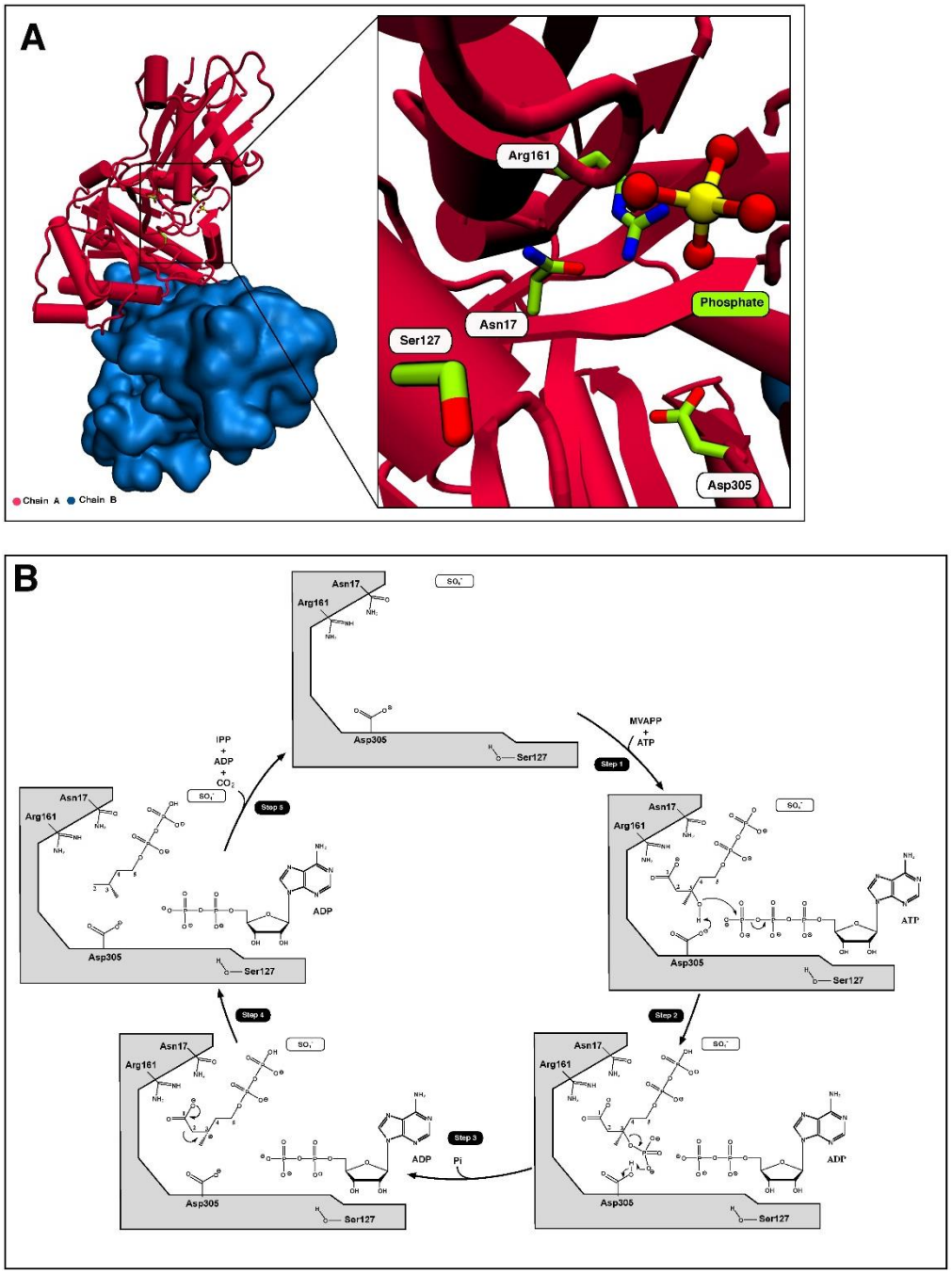


Figure 26 - A: Structure and active site of MDD (PDB code: 3D4J[259]). B: Proposed catalytic mechanism of MDD[259, 261-262].

Once the substrate, MVAPP, binds in the active site, it is suggested that it is positioned so that the terminal phosphate is located near the sulfate/phosphate anions that were detected in the X-ray structure (Figure 26-B step 1). The C3 hydroxyl group of

MVAPP stands in juxtaposition to the catalytic Asp305 and the ATP's gamma phosphoryl group is oriented to facilitate the gamma phosphoryl transfer to MVAPP, with the help of Ser127. The proposed general base catalyst, Asp305, is believed to deprotonate MVAPP's C3 hydroxyl in the first step of the catalytic process to facilitate the attack on the gamma phosphoryl of ATP (Figure 26-B step 2). From this reaction result the release of ADP and the formation of the 3-phosphoMVAPP intermediate[259]. The next step of the catalytic process involves the dissociation of phosphate, which generates a carbocation intermediate at carbon C3 (Figure 7-B step 3). The positively charged carbon at C3 creates an electron sink that accelerates the decarboxylation process with the help of Arg161. At the end of the reaction, carbon dioxide is generated concomitant with the formation of IPP (Figure 26-B step 4).

Inhibitors of MDD, in particular inhibitors of the bacterial enzyme form, are currently under development to be used as antimicrobial agents. Several substrate analogues, such as diphosphoglycolylproline (see PDB code: 4DU8[263]) and 6-fluoromevalonate diphosphate, were shown to be competitive inhibitors for the substrate in the bacterial enzyme.[263]

4.2.5. Isopentenyl-diphosphate Delta Isomerase

Isopentenyl-diphosphate delta isomerase (or isopentenyl pyrophosphate isomerase, IPP isomerase, EC 5.3.3.2) catalyzes the conversion of the relatively unreactive IPP to the more-reactive electrophile dimethylallyl pyrophosphate (DMAPP). Both reactants and products of this reaction are substrates for the successive reaction that results in the synthesis of farnesyl diphosphate (FPP) and, ultimately, cholesterol. This isomerization is therefore a key step in the biosynthesis of isoprenoids through the mevalonate pathway.

Crystallographic studies have shown that the active form of IPP isomerase is a monomer with alternating α-helices and β-sheets[264-265] (Figure 27-A). The active site of IPP isomerase is deeply buried within the enzyme and consists of a glutamic acid residue (Glu116) and a cysteine residue (Cys67) that interact with opposite sides of the IPP substrate, consistent with the antarafacial stereochemistry of isomerization. The enzyme also contains two metal sites: a first site occupied by Mn^{2+}, which is coordinated with three histidine and two glutamate residues; and a second site containing Mg^{2+}, which is coordinated with two pyrophosphate oxygens, the carbonyl group of the highly conserved residue Ala67, a carboxylate oxygen of Glu87 and two water molecules.

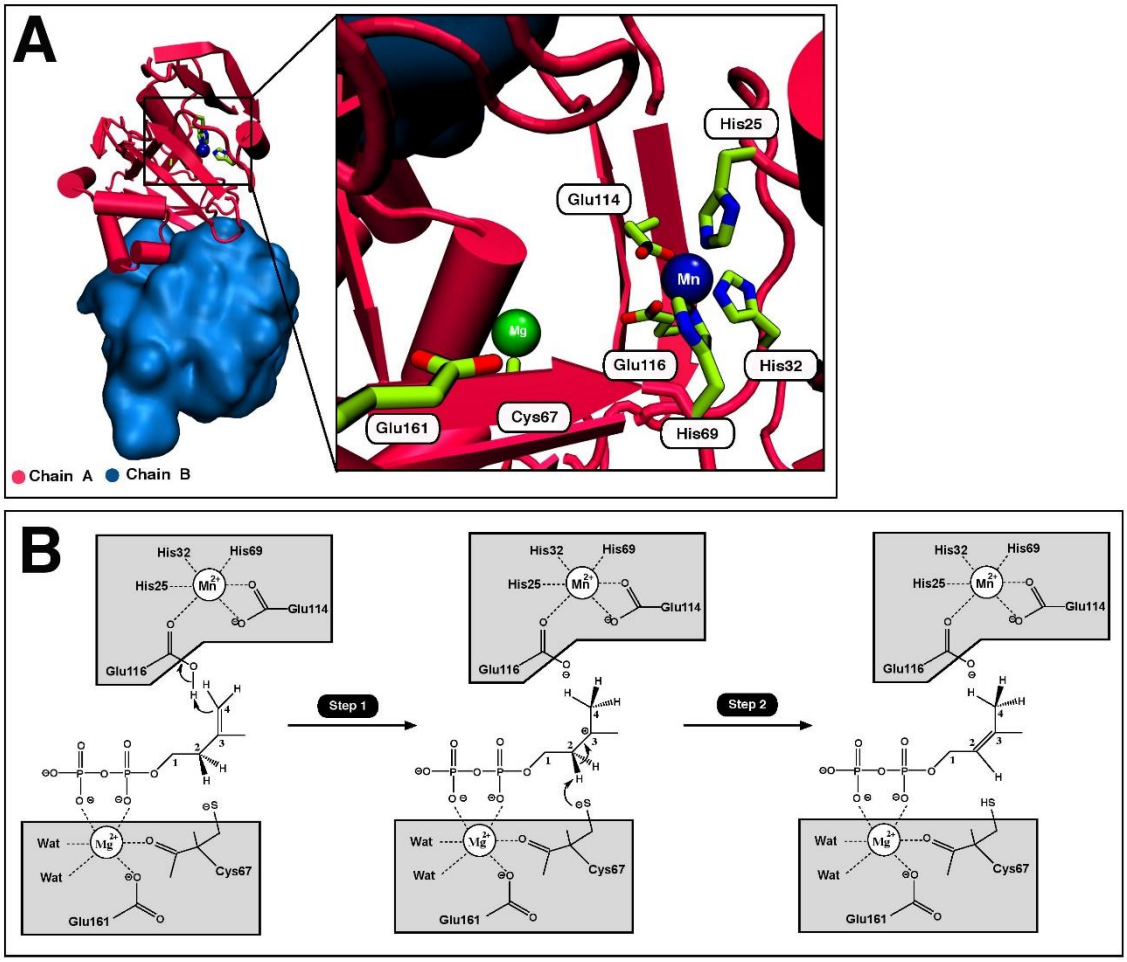


Figure 27 - A: Structure and active site of the enzyme isopentenyl pyrophosphate isomerase (pdb code: 1Q54[266]). B: Proposed catalytic mechanism for IPP isomerase[266].

The mechanism through which the enzyme undergoes the initial protonation step (Figure 27-B step 1) has not been conclusively established. Recent evidence suggests that Glu116 is involved in the protonating step of the C3-C4 double bond, forming a carbocation[266]. The thiolate of Cys67 then removes a proton from carbon C2, resulting in the isomerization of IPP into DMAPP (Figure 27-B step 2). During this process, the two metal centers have an active role in catalysis. The Mg^{2+} is required for the stabilization of the phosphate groups during the catalytic process, whereas the Mn^{2+} is actively involved in catalysis. It is also proposed that the reaction is facilitated by the conversion of the carboxyl group of Glu116 to a carboxylate, resulting in an electro-neutral Mn^{2+} with a bis-carboxylate complex. The water molecules present in the active site are also proposed to be involved in the catalysis, but the mechanism how this happens remains unknown[267-268].

4.2.6. Farnesyl Diphosphate Synthase

Farnesyl diphosphate synthase also known as farnesyl pyrophosphate synthase (FDPS, EC 2.5.1.10) is an Mg^{2+}-dependent homodimeric enzyme, localized in peroxisomes, which catalyzes a chain elongation reaction and controls the first branching point of the mevalonate pathway.

Chain elongation enzymes can be divided into two genetically different families depending on whether the stereochemistry of the newly formed double bond during each cycle of chain elongation is E or Z. FDPS is a member of the E-double bond family and catalyzes the sequential addition of IPP and DMAPP to form geranyl diphosphate (GPP) and then FPP.

FDPS is essential for the post-translational prenylation of all small GTPase proteins that play a crucial role in cell signaling, cell proliferation, and osteoclast-mediated bone resorption. Inhibition of FPP production can decrease the activity of mutated H-Ras, K-Ras, and N-Ras proteins that function as major drivers of tumor growth in many cancers[269-270]. Thus, the clinical benefits of human FDPS inhibition include both decrease of prenylation of mutated Ras proteins, leading to a decrease in cellular growth and/or survival, as well as alleviation of tumor-associated bone destruction via inhibition of osteoclast activity[271-272]. Additionally, blocking the catalytic activity of human FDPS impacts both the downstream and upstream levels of isoprenoids in the mevalonate pathway, leading to numerous cellular changes, including the inhibition of cholesterol biosynthesis[273-274]. Currently, FPPS inhibitors are mainly used in the treatment of a number of bone disorders, such as Paget's disease, hypercalcemia, metastatic osteolysis and osteoporosis[275].

The available X-ray structure for the human FDPS shows that the protein is a homo-dimer (Figure 28-A)[276-277]. Each subunit is folded in a single domain whose central feature is a core composed of 10 helices that surround a large deep cleft identified as the substrate-binding pocket. The active site is located in the bottom of the cleft and it comprises two conserved aspartate-rich motifs[278]. One includes the residues Asp117, Asp118, Ile119, Met120 and Asp121 and the other the residues Asp227, Asp258, Tyr259, Leu260 and Asp261. The aspartate-rich motifs are positioned so that the side chain carboxylate groups point into the cleft from opposite sides [279]. The Asp residues in the two aspartate-rich motifs, except the last one in the second motif, are important for catalysis, while the remaining residues of the second motif are essential for substrate binding. These results are consistent with the co-crystal structure of avian FPP in

complex with GPP and IPP[280]. The active site also lodges three magnesium cations, whose chelation with the substrate is required for enzyme activity.

Taking into account the crystallographic and kinetic evidence, two types of catalytic mechanisms have been proposed for FDPS: those in which condensation is initiated by heterolytic cleavage of the carbon−oxygen bond of the allylic pyrophosphate, yielding a cationic intermediate, and those where the formation of the C1-C4 bond between the two substrates and rupture of the C1 oxygen bond is simultaneous through a transition state (TS) with a carbocation character.

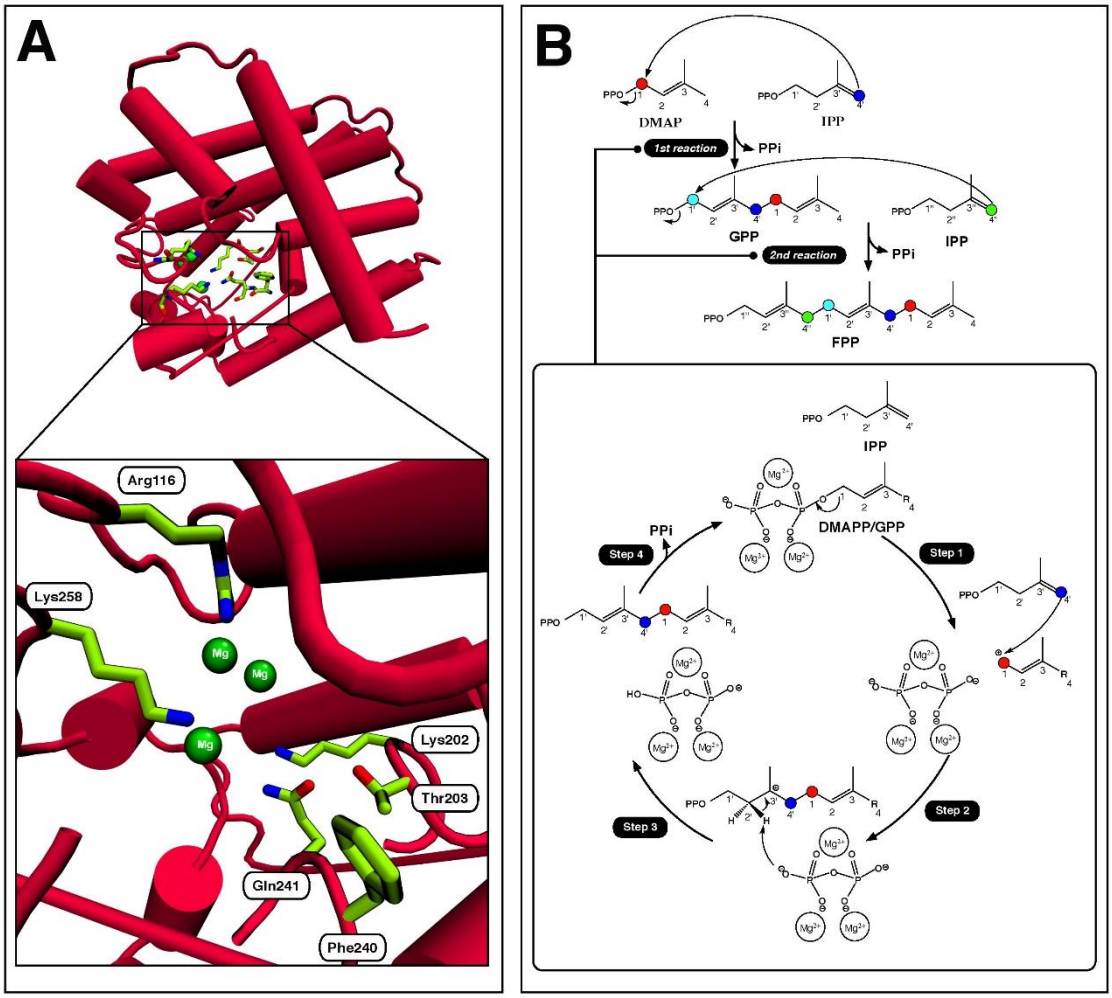
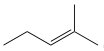


Figure 28 - A: Monomeric structure and active site of FDPS (PDB code: 1RQJ[281]). B: Catalytic mechanism of FDPS. R= Me or  , depending on whether FDPS catalyzes the first or second reaction[278,279].

Both mechanisms suggest the involvement of a dissociative electrophilic alkylation (Figure 28-B). During the reaction, the bond between carbon C1 and its neighboring

oxygen in the diphosphate (in DMAPP or GPP) is cleaved to generate a resonance-stabilized allylic cation, which then alkylates the double bond in IPP to produce a tertiary carbocation. Whether this occurs in two independent steps or in a concerted manner is still a matter of some controversy, but recent QM/MM studies suggest that the latter type of mechanism is more favorable[282]. It is also suggested that the main chain carbonyl oxygen of Lys202, and the side chain oxygens of Thr203 and Gln241 are important to stabilize the allylic cation transition state (Figure 9-A).

The next step involves the elimination of a proton from carbon C2 of the IPP unit to give rise to a new E-double bond between carbons C2 and C3. Hosfield et al, suggest that the non-metal-ligated pyrophosphate oxygen is the catalytic base that deprotonates the condensed intermediate to generate the C5-extended isoprenoid reaction product, and that Arg116 and Lys258 might be important for such reaction[281]. The new allylic diphosphate is then generated and it is one isoprene unit longer than the initial substrate. The role of the magnesium cations present in the active site is believed to be the stabilization of the pyrophosphate groups and, at the same time, the enhancement of the rate of the reaction.

4.2.7. Squalene Synthase

Squalene synthase or farnesyl-diphosphate:farnesyl-diphosphate farnesyl transferase (SQS, EC 2.5.1.21) catalyzes an unusual head to head reductive dimerization of two molecules of FPP to form squalene. This enzyme is of particular importance as it is the first enzyme in the pathway responsible for the production of a metabolite that is solely committed to cholesterol synthesis.

SQS is located in the membrane of the endoplasmic reticulum. It is anchored to it by a short C-terminal membrane-spanning domain, whereas the N-terminal catalytic domain of the enzyme protrudes into the cytosol.

Mammalian forms of SQS are approximately 47 kDa and consist of approximately 416 amino acids. The crystal structure of human SQS was determined in 2000[283], and revealed that the protein was composed entirely by α-helices. The active site of this enzyme is located in a large channel that is found in the middle of the protein. One end of the channel is open to the cytosol, from where the substrates are obtained (soluble allylic compound containing 15 carbon atoms). The other end forms a hydrophobic pocket that allows the product of the catalytic process to reach the membrane environment (squalene - an insoluble compound with 30 carbon atoms) (Figure 10-A).

The detailed mechanism by which SQS operates has not been fully resolved yet. The uniqueness of the head-to-head coupling of two FPP molecules to form squalene via a stable cyclopropylcarbinyl diphosphate intermediate has elicited much mechanistic speculation over the years[284-285]. In order to shed some light on the catalytic mechanism of SQS, several site-directed mutagenesis experiments on diverse amino acid residues of the SQS active site were conducted. The results showed that mutations of Phe288 prevent the enzyme from catalysing the second reaction, in which the cyclopropyl intermediate is rearranged to produce squalene[286]. On the other hand, mutations at Tyr171 hinder the enzyme inactive. This result suggests that Tyr171 is required for the first half reaction: the ciclopropanation. Because tyrosine is an aromatic residue and has a hydroxyl group, a proposal was created in which Tyr171 loses a proton to the leaving pyrophosphate group. This would turn Tyr171 more negative, which in turn would stabilize the carbocation intermediate that is generated during the reaction. In 2000 and with the release of the SQS structure[283], this hypothesis gathered more consensus, as Tyr171 was found to be in the binding pocket. Also, likely to play a key role in catalysis are the conserved Tyr73 and Gln212. In addition, three magnesium ions are important for the catalysis and might have a major role in the stabilization and orientation of the two FPP molecules in the active site of SQS.

In 2010, the crystal structure of a very similar enzyme, dehydrosqualene synthase (crtM), was crystallized with a substrate analogue, farnesyl thiopyrophosphate (PDB code 3W7F[287]) and the intermediate, PSDP (PDB code: 3NPR[288]). Superimposition of the substrate and intermediate in the SQS structure revealed that one of the substrate molecules remains in the same position during the first half reaction, and keeps the pyrophosphate moiety (Figure 29-A).

Based on the available data, it is suggested that the reaction catalyzed by SQS proceeds in two distinct steps, both of which involve the formation of carbocationic reaction intermediates (Figure 29-B). In the first half-reaction, two identical molecules of FPP bind in distinct regions of the active site of SQS. The pyrophosphate group of one of the FPP (normally designated as donor FPP) is then cleaved and an allylic carbocation intermediate is generated concomitant with the release of pyrophosphate (Figure 29-B step 1). The carbocation promptly reacts with the C-2,3 double bond of the acceptor FPP in a 1',2,3 prenyl transferase reaction that is accompanied by the loss of a proton (Figure 10-B step 2). The product of this condensation is presqualene pyrophosphate (PSPP), a stable cyclopropylcarbinyl pyrophosphate intermediate that remains associated with SQS for the second reaction (Figure 29-B step 3).

In the second half-reaction of SQS, PSPP loses the pyrophosphate group and the resulting cyclopropylcarbinyl carbocation undergoes a ring opening and reduction by NADPH (through a hydride transference) to the linear and final product, squalene (Figure 10-B step 4,5 and 6). SQS, then releases squalene into the membrane of the endoplasmic reticulum (Figure 29-B step 7).

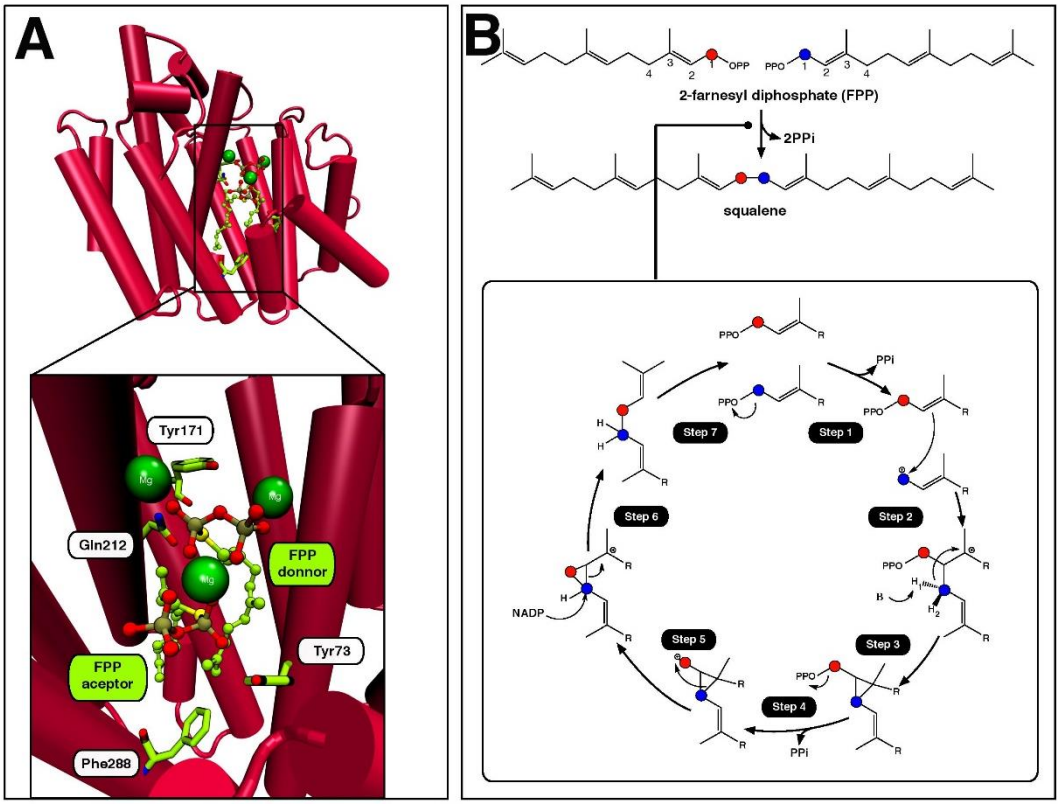


Figure 29 - A: Structure and active site of squalene synthase (SQS). The image of the SQS active site was built through the superimposition of two X-ray structures: 3W7F (enzyme analogous to SQS, carotenoid dehydrosqualene synthase, which catalyses a similar reaction)[287] and 1EZF[283] (Human SQS). The amino acid residues represented on figure A are from the Human SQS. Only the FPP was retrieved from the PDB code 3W7F B: Proposed catalytic mechanism for SQS[283, 287].

From a metabolic perspective, SQS has a pivotal role as it takes FPP from the more general mevalonate pathway, to commit it into the cholesterol synthesis[289]. This raised interest in using SQS as a drug target for the treatment of hypercholesterolemia. Currently, strong inhibitors against SQS have been found, such as Zaragozic acid A[290-291], Lapaquistat (recently discontinued from clinical development)[292] and DF-461[293]. Several authors suggest that the use of SQS in the treatment of hypercholesterolemia could have fewer side effects than the classical statin treatments, which inhibit the whole mevalonate pathway by targeting HMG-COA-R. At the moment this is still speculative,

since only one inhibitor of SQS, lapaquistat, went through clinical trials, but failed due to potential hepatic safety issues[294].

SQS was also pointed out as an interesting chemotherapeutic target against Trypanosoma cruzi and Leishmania parasites, which are responsible for large parasitic disease loads and socioeconomic losses, particularly in developing countries[295]. Specific chemotherapy of the diseases caused by these protozoa remains unsatisfactory, as currently available drugs have limited activity, as well as frequent toxic side effects and rising drug resistance. However, these parasites have a strict requirement for specific endogenous sterols (ergosterol and analogs) for survival and growth and cannot use the abundant supply of cholesterol present in their mammalian host[296-297]. Further studies led to the discovery of E5700 and ER-119884, two novel quinuclidine SQS inhibitors under development as cholesterol and triglyceride lowering agents in humans by Eisai Company, which have very potent anti-*T. cruzi* activity *in vitro*. One of them (E5700) was able to provide full protection against death, and it completely arrested development of parasitaemia in a murine model of acute disease when given orally. This was the first report of an orally-active SQS inhibitor as an anti-infective agent[298-299]. Although these compounds and other aryl-quinuclidines are also potent inhibitors of mammalian SQS, their selective antiparasitic activity in vitro and in vivo might be explained by the capacity of a host's cells to compensate for the blockage of de novo cholesterol synthesis by up-regulating the expression of LDL receptors and capturing this sterol from the growth medium or serum[300-301]. In contrast, there is no way for the parasite to compensate in this manner for the quinuclidine-induced blockage of ergosterol biosynthesis, since there are no appreciable amounts of ergosterol in host cells or growth media.

4.2.8. Squalene Monooxygenase

Squalene monooxygenase (SM, EC: 1.14.13.132), also known as squalene epoxidase, is a 64-kDa flavin adenine dinucleotide (FAD)-containing enzyme that catalyses the first oxygenation step in the cholesterol synthesis. The reaction requires NADPH and molecular oxygen to oxidize squalene, to 2,3-oxidosqualene (squalene epoxide).

SM is bound to the endoplasmic reticulum of eukaryotic cells and belongs to the flavoprotein monooxygenase family, which catalyzes a wide variety of oxidative reactions[214, 302]. However, there are several differences between squalene monooxygenase and other known flavin monooxygenases. Firstly, SM catalyzes epoxidation but not hydroxylation of substrates. Secondly, it does not contain cytochrome

P450 as a prosthetic group and does not bind NADPH directly. Instead, the electrons that are required for the reaction are passed from NADPH, via cytochrome P450 reductase, to the loosely bound FAD group of squalene monooxygenase. These differences turn SM into an attractive target to treat hypercholesterolemia, while not disturbing other flavin monooxygenases dependent processes.

SM was first identified in the early 70s, but studies on the human enzyme began only in the 90s. The mammalian SM showed limited sequence similarity to that found in yeast. Interestingly, it has also been shown that there are substantial differences between the mammalian enzymes from different species[303-304], revealing that they might be specific to each specie. Unfortunately, structural details of these enzymes from their X-ray diffraction studies are still not available, which impedes further analysis.

The reaction catalyzed by squalene monooxygenase is highly specific[305]. It involves insertion of an oxygen atom across a carbon-carbon double bond to form an epoxide. Flavoprotein monooxygenases accomplish this oxygenation by forming a flavin hydroperoxide at the enzyme active site, which then transfers the terminal oxygen atom of the hydroperoxide (OH) to the substrate[306] . The remaining "hydroxyflavin" then reoxidizes, with the release of a water molecule and the concomitant formation of 2,3-oxidosqualene. Squalene monooxygenase presumably utilizes this same mechanism, but differs from other known flavin monooxygenases in that the oxygen is inserted as an epoxide rather than as a hydroxyl group. Another thing that differentiates these enzymes is that squalene monooxygenase contains a loosely bound FAD flavin and obtains electrons from NADPH-cytochrome P450 reductase, rather than binding the nicotinamide cofactor NADPH directly[307].

Although there is a lack of information regarding the structure and biochemistry of SM, this enzyme has been extensively exploited for antifungal drug development[308-309]. Currently, several fungal SM inhibitors, such as terbinafine, naftifine and numerous related compounds, are currently on the market or under investigation. Some of the best compounds are Terbinafine (Lamisil®) and Naftifine (Naftin®), which show good specificity for the fungal enzyme without inhibiting human squalene monooxygenase[310]. Both are fungicidal, interfering with cell membrane synthesis and preventing growth.

Since this enzyme catalyses the second committed (and likely rate-limiting) step in cholesterol biosynthesis, it has also become an attractive pharmacotherapeutic target in the treatment of hypercholesterolemia and resultant cardiovascular diseases. Several preclinical studies suggest an effective hypocholesterolemic activity, comparable or even better than that which can be achieved by HMG-CoA-R inhibition.

The most effective inhibitor of mammalian SM known to date is NB-598, developed at Banyu Pharmaceutical Co[311]. This fungal-derived natural compound is a competitive inhibitor of SM in human HepG2 cells and effectively reduces serum cholesterol in dogs with no apparent adverse effects[312]. In a comparative study in dogs, NB-598 was more potent than the HMG-CoA-R inhibitor simvastatin, decreasing total and serum LDL cholesterol. NB-598 also decreased triacylglycerol levels, an effect not observed with the HMG-CoA-R simvastatin[312]. No studies have yet been reported in man.

4.2.9. 2,3-Oxidosqualene Cyclase-lanosterol Synthase

2,3-Oxidosqualene cyclase-lanosterol synthase (OSC, EC 5.4.99.7) is a 78 kDa membrane-bound enzyme that catalyses the conversion of the acyclic compound 2,3-oxidosqualene (OS) to the cyclic lanosterol. Both oxidosqualene and lanosterol are mostly hydrocarbons, and thus are not very soluble in water. The enzyme solves this problem by sticking to the membrane in peroxisomes. It then can pull oxidosqualene directly out of the membrane, and release lanosterol back there. The reaction catalysed by this enzyme is considered to be one of the most complex reactions that are catalysed in the enzymatic world, as it involves a sequence of cyclization and 1,2-group rearrangements of high energy carbocations to lanosterol, with complete structural and stereochemical control of seven chiral centers[313].

OSC is almost exclusively found in eukaryotes, apart from a few prokaryotes that can produce this enzyme[314]. Human OSC is a monomeric enzyme that it is located in the membrane of the endoplasmic reticulum[315]. OSC is composed by two α-helix barrel domains (domain 1 and domain 2) connected by loops and three smaller β-sheet structures. The active site cavity is located in the middle of the protein, between domains 1 and 2 (Figure 30-A)[313]. Domain 2 contains a region inserted in the membrane and forms a tunnel that allows the substrate to enter the enzyme and reach the active site cavity. Although the architecture of the channel is clear in the available X-ray structure, its role in enzyme-subtract interaction is still far from being understood.

The catalytic mechanism of OSC has been a subject of numerous studies in the past four decades[313, 316]. In spite of these efforts, much remains to be learned about how OSC mediates the individual cyclization and rearrangement steps. Currently, there is still a great speculation about the mechanism of OSC and several contradictory proposals are present in the literature. In general, all the authors agree that the catalytic mechanism of OSC can be divided in three main stages. The mechanism starts with the adoption of a pre-organized conformation by 2,3-monoepoxysqualene, followed by protonation of the

epoxide ring. A cascade of reactions is then triggered, resulting in an array of ring-forming reactions followed by a series of 1,2-hydride and 1,2-methyl group shifts and a final deprotonation step, from which results lanosterol (Figure 11-B). All reactions proceed with precise stereo- and regiospecificity required for the formation of the multi C-C bonds.

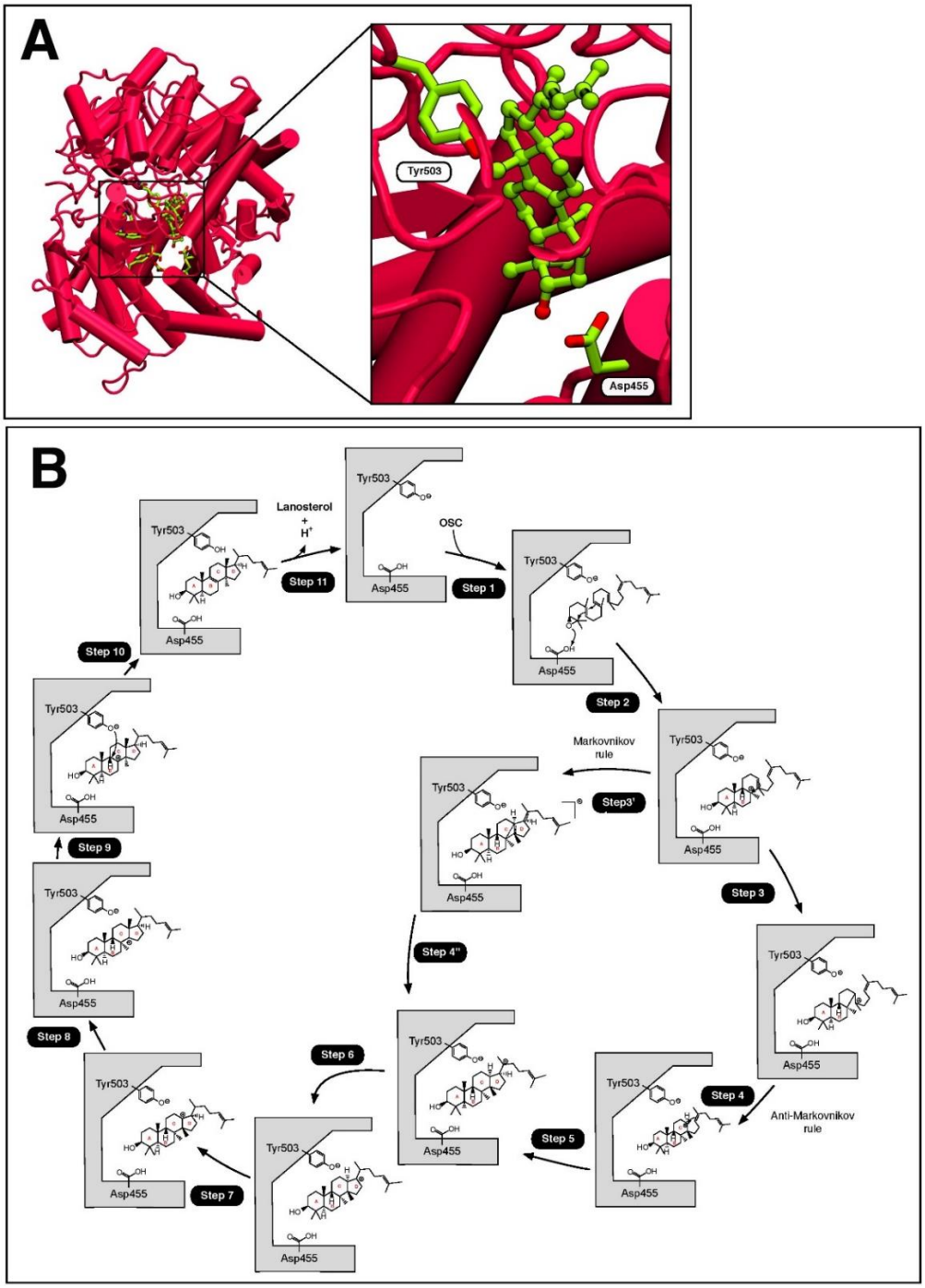


Figure 30 - A: Structure of the enzyme and of the active site of OSC (pdb code 1w6k). B: catalytic mechanism catalyzed by OSC.

The first stage of the reaction is assisted by Asp455, which protonates the epoxide of the substrate (Figure 30-B step 1). The acidic character of this residue is enhanced by the neighboring Cys465 and Cys533 with which it interacts through two hydrogen bonds. The second stage of the mechanism was initially thought to involve the sequential formation of the four rings (named A, B, C and D) in a concerted manner[316-324]. Corey et al proposed that the formation of the A-ring should be concomitant with the protonation of the epoxide, in a concerted SN2 substitution like process. This proposal was confirmed very recently by QM/MM calculations[325]. The same calculations have also suggested that the formation of the B-ring is concerted with this step, and that the three concomitant reactions are the rate-limiting step of the full catalytic process (Figure 30-B step 2).

Corey and co-workers also proposed that formation of the rings C and D involved discrete carbocation intermediates. In particular, they showed that the formation of ring C involves a cyclopentylcarbinyl carbocation intermediate that subsequently undergoes a ring expansion to form the six-membered C-ring. It is also proposed that the closure of the D-ring is concerted with the formation of the C-ring[326-327] (Figure 30-B step 3', 4'). The only drawback of this proposal is the formation of ring C, which does not follow the Markovnikov rule and, therefore, does not ensure the formation of a stable carbocation during the addition reactions. In 2002, Hess presented another hypothesis for C and D ring formation[317]. The double bond between carbon 18 and 19 of squalene might be involved anchimerically in the cyclopentylcarbinylcyclohexyl ring expansion, meaning that the expansion of the C-ring and formation of the D-ring are concerted. This mechanism avoids the formation of the stable cyclohexyl carbocation intermediate as well as the violation of Markovnikov's rule (Figure 30-B step 3, 4, 5). Although this mechanism is more favorable than the previous one, there is still no certainty about which one actually occurs. The only detail that seems clear is that the formation of rings C and D is controlled by a set of intramolecular forces (cation stabilization and hydrogen-bonding) that are imposed by some active site residues to the substrate and are located along the active site tunnel. Mutagenesis studies have revealed that Tyr704, Tyr707, Tyr587 and His234 have an important role in this context[328].

Once the formation of the D-ring is accomplished, a cation intermediate is obtained, which undergoes a series of 1,2-methyl and hydride shifts (Figure 30-B step 6, 7, 8, 9). The last step of the catalytic cycle involves a proton abstraction which is catalyzed by Tyr503 (Figure 30-B step 10). In the end of this step the product of the reaction, lanosterol, is obtained and the enzyme is ready for a new turnover (Figure 30-B step 11).

The interest in OSC has grown in recent years, due mostly to its late position in the cholesterol biosynthesis, which means it can be an optimal target to fight

hypercholesterolemia. This interest is partially motivated by the adverse effects of statin that inhibit HMG-CoA-R and may cause myopathy. However, until the present date, no OSC inhibitor has been approved as a clinical drug for the treatment of hypercholesterolemia. However, some OSC inhibitors are already been described and co-crystalized in the active center of a close homologue of OSC. One in particular, named Ro 48-8071, is one of the most popular among OSC inhibitors, being co-crystalized in both human OSC enzyme (PDB code 1W6J[313]) and in the homologous SHC enzyme (PDB code 3SQC[329]). It acts on human OSC as competitive inhibitor, preventing the binding of the oxidosqualene.

4.2.10. From Lanosterol to Cholesterol

The conversion of lanosterol into cholesterol is a very complex and multistep pathway, which involves several enzymes. After squalene is transformed into lanosterol, this molecule can follow two different routes, both of which ends with a cholesterol molecule. They are referred to as the Bloch pathway and the Kandutsch-Russell pathway. The difference between them is that the first uses Δ^{24}–unsaturated sterols while in the second the intermediates have their side chain saturated.

In order to be transformed into a cholesterol molecule, lanosterol has to undergo a series enzymatic reactions (Figure 31). First, C-14 undergoes a two-step demethylation process, which is catalyzed by the enzymes C14α-demethylase and Δ^{14}-reductase. This is followed by two subsequent demethylations at C-4, mediated by C4-demethylase. The next step is the isomerization of the double bond at Δ^{8} to Δ^{7}, a reaction catalyzed by Δ^{8}-Δ^{7}-isomerase. Subsequently, a desaturation occurs between C-5 and C-6 (catalyzed by Δ^{5}-desaturase) followed by the reduction of two double bonds, the first at Δ^{7} and the second at Δ^{24}. The enzyme Δ^{24}–reductase can convert any intermediate on the Bloch pathway into its unsaturated counterpart, although it as different affinities for them, and the conversion will shift to the Kandutsch-Russell pathway. However, after this route is chosen, it cannot revert to the previous one. There is a great lack of structural and mechanistic information about this final stage of the cholesterol biosynthetic pathway and for that reason it was considered premature to explore it in the present review.

4.3. Conclusions and Future Perspectives

The complexities of the structure and the biosynthesis of cholesterol have spanned decades of research and gathered several Nobel prizes. It is perhaps the only molecule

whose studies have paved the way for the discovery of several and important biomedical benefits. After one century of research, significant advances have taken place, and our current knowledge regarding the enzymes involved in steroid hormone biosynthesis has increased substantially. This means that we are now closer to deciphering the full mechanistic puzzle behind cholesterol biosynthesis.

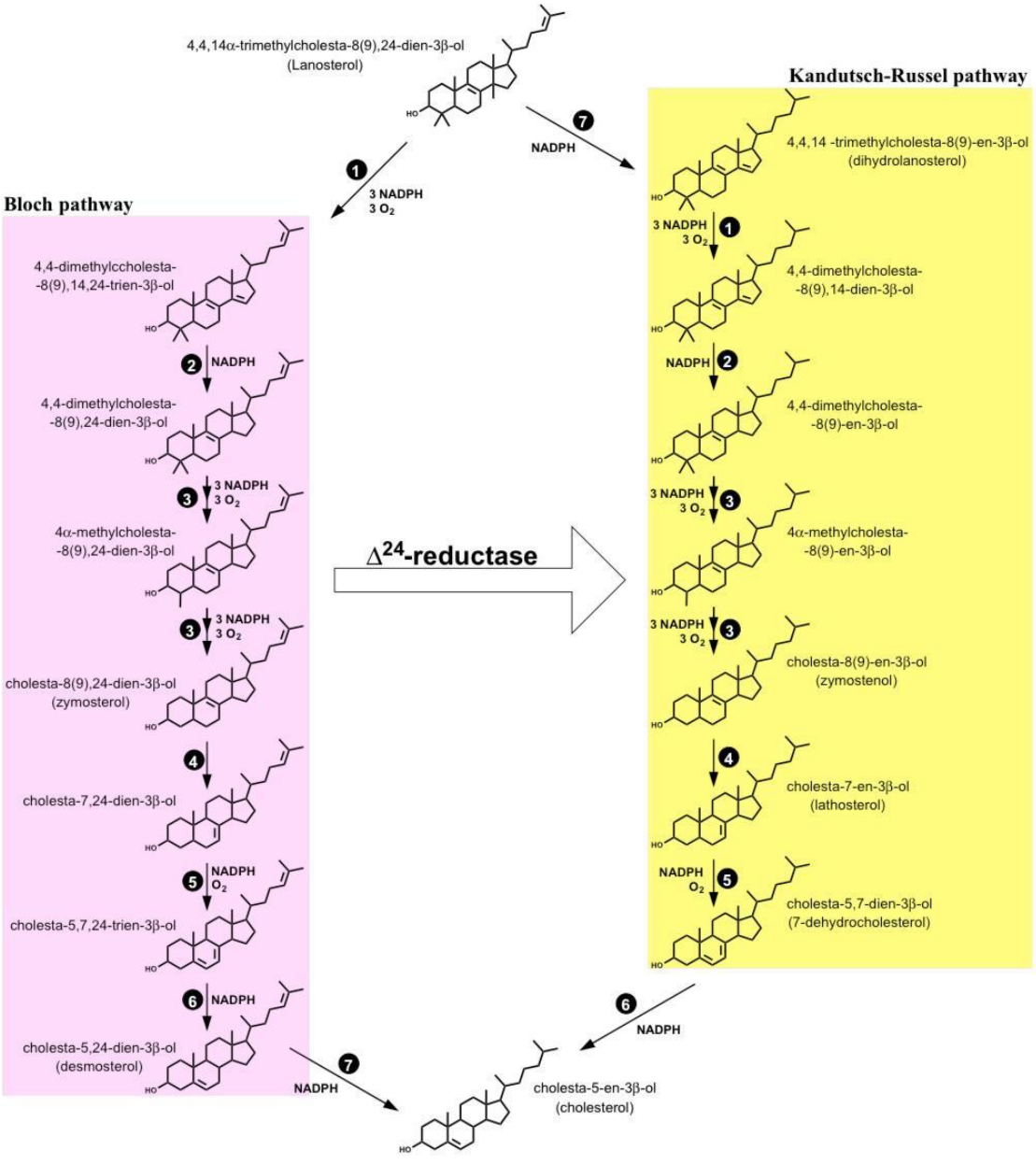


Figure 31 - Enzymes involved in the catalysis of cholesterol from lanosterol.

Currently, all the enzymes involved in the biosynthesis of cholesterol have already been identified and the three-dimensional structures of many of them have been deciphered and intensively studied. In this review, all the mechanistic knowledge

regarding the majority of the enzymes involved in the biosynthesis of the cholesterol molecule, starting from acetyl-CoA, was explored and discussed. The structure of each enzyme involved in this process and a detailed description of their active site was scrutinized. Particular attention was given to the catalytic mechanism of each enzyme, and the chemistry behind the transformations from the reactants into the products was carefully described. The chemistry of these transformations is extremely challenging, involving, in the many cases, molecules that are very hydrophobic. Such a rich chemistry is only possible due to the involvement of remarkable enzymes that allow the binding of more than one substrate in their active sites, stabilize uncommon intermediates, manage the full stereochemical control of chiral centers, among others. The mechanisms through which these enzymes achieve the enormous rate enhancements and exquisite specificity are also diverse. For instance, the first three enzymes involved in the cholesterol biosynthesis, ACAT, HMG-CoA-S, and HMG-CoA-R, follow a ping-pong type of mechanism involving the formation of tetrahedral intermediates with active site residues. The next three enzymes, MK, PMK, and MDD are ATP dependent and involve the exchange of phosphoryl groups. The enzymes FDPS and SQS follow a sequential addition of substrates. SM catalyzes the oxygenation of a squalene and OSC isomerizes oxidosqualene to lanosterol.

In the quest of unraveling the catalytic mechanism of the enzymes involved in the cholesterol pathway, the theoretical and computational methods are gaining an important role, since they can predict and tackle the formation of reaction intermediates that are difficult to detect by experimental means, as it has been observed in other biological systems[330-335]. Some of these proposals were validated later on by experimental means, while others were used to explore new mechanistic variants[317, 325].

From all the described enzymes in the cholesterol biosynthesis, HMG-CoA-R, SQS and OSC are perhaps the most important, from a pharmacological point of view. HMG-CoA-R is the target of statins, important drugs that lower blood cholesterol levels and treat cardiovascular diseases. However, this enzyme acts very early in the cholesterol synthesis pathway, with nearly 20 subsequent enzymes needed to produce cholesterol. SM and SQS are enzymes which are downstream from HMG-CoA-R and located in the final branching point of the cholesterol biosynthesis. The inhibitors for these enzymes are not so well develop as the ones for HMG-CoA-R, but currently available studies show very promising results in reducing cholesterol. In fact, preclinical studies suggest that the inhibition of SM has hypocholesterolemic activity comparable, or even better, than that obtained with HMG-CoA-R inhibition. Studies involving OSC also show direct decrease in lanosterol formation and an inherent decrease in HMG-CoA reductase activity. The

great advantage of these inhibitors in relation to the ones that target HMG-CoA-R is the absence of myopathic side effects, one of the major drawbacks of statins.

It is also worth mentioning that there are currently other efficient methods that can be used to treat hypercholesterolemia. For example, the bile acid binding sequestrant cholestyramine has been used for several decades to treat hypercholesterolemia, even before statins became available on the market[336]. The cholesterol absorption inhibitor ezetimibe can also be used to the same end, alone or combined with statins[337-338]. Recently, it was also found that inhibitors of proprotein convertase subtilisin/kexin9 (PCSK9) monoclonal antibody can be used effectively to treat hypercholesterolemia[339-340]. The mechanism of action of these drugs is still poorly understood but it is believed that they might not interact directly with the enzymes involved in cholesterol biosynthesis.

All of these new compounds and methods constitute a great promise and may become the next generation of drugs that will be used to reduce low-density lipoprotein cholesterol levels, possibly in a future not far from today.

CHAPTER 5

UNRAVELING THE ENIGMATIC MECHANISM OF L-ASPARAGINASE II WITH QM/QM CALCULATIONS

In this paper, we have studied the catalytic mechanism of L-asparaginase II computationally. The reaction mechanism was investigated using the ONIOM methodology. For the geometry optimization we used the B3LYP/6-31G(d):AM1 level of theory and for the single points M06-2X/6-311++G(2d,2p):M06-2X/6-31G(d) level of theory. It was demonstrated that the full mechanism involves three sequential steps and requires the nucleophilic attack of a water molecule on the substrate prior to the release of ammonia. There are three rate limiting states, which are the reactants, the first transition state and the last transition state. The energetic span is 20.2 kcal/mol, which is consistent with the experimental value of 16 kcal/mol. The full reaction is almost thermoneutral. The proposed catalytic mechanism involves two catalytic triads that play different roles in the reaction. The first triad, Thr12-Lys162-Asp90 acts by deprotonating a water molecule that subsequently binds to the substrate. The second triad, Thr12-Ty25-Glu283, acts by stabilizing the tetrahedral intermediate that is formed after the nucleophilic attack of the water molecule to the substrate. We have shown that a well-known Thr12-substrate covalent intermediate is not formed in the wild-type mechanism, even though our results suggest that its formation is expectable in the Thr89Val mutant. These results have provided a new understanding of the catalytic mechanism of L-asparaginases that is in agreement with the available experimental data, even though it is different from all earlier proposals. This is of particular importance since this enzyme is currently used as a chemotherapeutic drug against several types of cancer and in the food industry to control the levels of acrylamide in food.

Adapted from reference [341]

For this paper Diana Gesto preformed all the calculations and analyzed the results, wrote the entire preliminary draft of the manuscript which was then reviewed by all co-authors.

5.1. Introduction

Recent studies show that the decrease of the concentration of highly expressed amino acids in tumors can retard or even stop tumor growth without affecting the metabolism of normal cells.[342-343] The starvation of cancer cells through amino acid deprivation has thus become an encouraging strategy in cancer therapy. However, restricted diet is not enough to control the concentration of these amino acids in the blood serum. Therefore, the administration of enzymes specifically addressed to metabolize these amino acids is the method currently used. One of the therapies recently approved by the FDA to decrease amino acid blood pools is the administration of the enzyme L-asparaginase.[342]

L-asparaginase (L-asparaginase amidohydrolase, EC 3.5.1.1) is an enzyme that hydrolyzes L-asparagine to L-aspartate, with the release of one ammonia molecule (Figure 32). It is currently used in both cancer therapy and food industry. This enzyme is an important chemotherapeutic drug approved by the FDA and with activity against several types of cancer, such as acute lymphoblastic leukemia[344-345], lymphosarcoma[345] and a few sub-types of non-Hodgkin's lymphoma.[346] The mechanism through which L-asparaginase destroys cancer cells is related to the fact that, in certain types of tumor, the production of asparagine synthase is limited. These cells are, therefore, incapable of producing enough amounts of asparagine to support their rapid growth, which promotes their dependence on external sources of this amino acid. The treatment with L-asparaginase reduces the levels of asparagine in the blood stream and, since normal cells are capable of producing enough quantity of this amino acid to survive, only cancer cells will be selectively affected.[347] This means that L-asparaginase can be efficiently used to control the growth of the tumor, and eventually to destroy it, while leaving the normal cells unharmed. Although L-asparaginase has been identified in many organisms, only enzymes with bacterial origin are used in cancer therapy, and from these, only type II L-asparaginases have anticancer activity. This is due to the fact that type II bacterial L-asparaginase has greater affinity to asparagine ($K_m = 1.15x10^{-5}$) when compared with type I ($K_m = 3.5x10^{-3}$).[348] Since only L-asparaginase II shows anti-cancer properties, our study was mainly focused on this enzyme[342-343]. In food industry, L-asparaginase is also being used to reduce the formation of acrylamide from starchy foods.

Despite all these current uses of L-asparaginase in medicinal and industrial applications, the reaction mechanism of this enzyme is still not known to its full extent. Taking this into account, we have studied the mechanism by computational means.

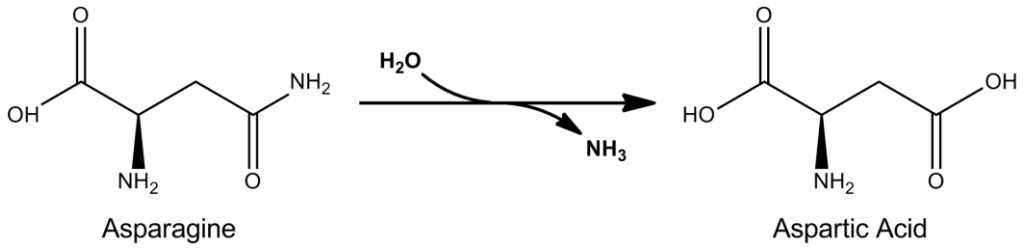


Figure 32. Reaction catalysed by L-Asparaginase.

Research on L-asparaginase and related amidohydrolases has been going on for over 40 years, but the correct understanding of the catalytic mechanism remains enclosed on the available X-ray structures. In this period of time, the detailed enzymological study of L-asparaginases has been focused mainly on *Escherichia coli* asparaginase (EcA). However, several studies have shown that the enzymatic mechanisms of all type II amidohydrolases are very similar.

The first crystallographic model of this enzyme was that of *Acinetobacter gluataminasificans* glutaminase-asparaginase, reported in 1988.[349] Several crystal structures have been reported since then, some containing only the enzyme [350-351], while others are complexes of the enzyme with other compounds such as the substrate l-asparagine, the product l-aspartate [352-353], the alternative products D-aspartate, L-glutamate and L-succinic acid [354], or even suicide inhibitors such as the L and D-stereoisomers of 6-diazo-5-oxy-norleucine.[355]

Detailed analysis of the active sites of bacterial L-asparaginase II of *E.coli*, *Erwinia carotovora* and *Erwinia chrysanthemi* showed a remarkable structural conservation among L-asparaginases. The binding pocket of these enzymes involves residues from both subunits of the intimate dimer, i.e. Thr12A, Tyr25A, Ser58A, Gln59A, Thr89A, Asp90A and Lys162A from one subunit (chain A), and Asn248C and Glu283C from the other subunit (chain C), that are interconnected by strong hydrogen bonds. In addition, one water molecule is structurally conserved and is part of a well-defined hydrogen-bond network (WAT1355 – PDB code 3ECA). This indicates that it is an integral part of the active-site architecture and may play a significant role in the catalytic properties of L-asparaginase.

The position of the amino acid residues in the active site suggests several possible pathways for the catalysis, but the almost symmetric location of two threonine residues, Thr12 and Thr89, above and below carbon C2 (please refer to Figure 35 for numbering) of the substrate suggests that one of them must be directly involved in the reaction.[353]

In order to decipher the involvement of these two residues in the reaction, Harms, Derst and Palm mutated these residues and evaluated the activity of the enzyme. In 1991 and 1992, Harms[356] and Derst[357] showed that the activity of the enzyme significantly decreased when Thr12 was mutated by an alanine residue (specific activity of the mutated enzyme was less than 0.01 U/mg, against 150 U/mg for the wild type enzyme).[356] However, when it was substituted by a serine residue, no change in the enzymatic activity was observed.[357] In 1996, Palm et al. revealed that when Thr89 was mutated for a valine, the enzymatic turnover was precluded due to the formation of an acyl intermediate between Thr12 and the substrate.[358] These results seem to indicate that Thr12 plays an important role in the reaction, while Thr89 may only be needed on a subsequent step.

Based on these data, many authors propose that in the wild-type enzyme, the formation of an acyl-intermediate should exist and would involve the participation of Thr12. Nevertheless, the results are somewhat contradictory since, when Thr12 is mutated by an alanine, the enzyme remains active, but it is much less efficient (less than 0.01% of the specific activity of the wild-type enzyme).[356] In addition, substrate analogues have also been found covalently bound to other active site residues, such as Ser9, which is not even a conserved residue in the L-asparaginase structures.[359]

Another site-directed mutagenesis study evaluated the role of Tyr25 in the reaction. This residue, together with Thr12 and Thr89, is also conserved in all L-asparaginases and interacts very closely with Thr12 through a hydrogen bond. The idea here was to replace Tyr25 with a phenylalanine, in order to evaluate the role of the hydroxyl group of this amino acid in the reaction. The final results showed that the mutation does not have a high impact in the activity of the enzyme, and only moderately affects the K_m value. Other Tyr25 mutants (Tyr-25-Ala, Tyr-25-Gly) presented similar results, suggesting that Tyr25 may not be directly involved in the reaction, but required for the activity of the enzyme, or perhaps it might be important for the specific recognition of the native substrate.[360]

In spite of the contradictory information that all these results have generated, the currently accepted mechanism of L-asparaginases proceeds via a covalently bound enzyme intermediate as depicted in Figure 33. This implies the initial nucleophilic attack of an active site amino acid (Thr12) to the substrate (asparagine), the subsequent release of ammonia, and the formation of an acyl-enzyme intermediate, in which the substrate becomes covalently bound to the enzyme. This intermediate is then attacked by a second nucleophile, usually water, resulting in the hydrolysis of the acyl-enzyme intermediate yielding the acidic product (glutamate) and free enzyme. In this process,

Lys162 has been proposed to have an active role and act as an acid and base in a proton buffer process with Thr89.

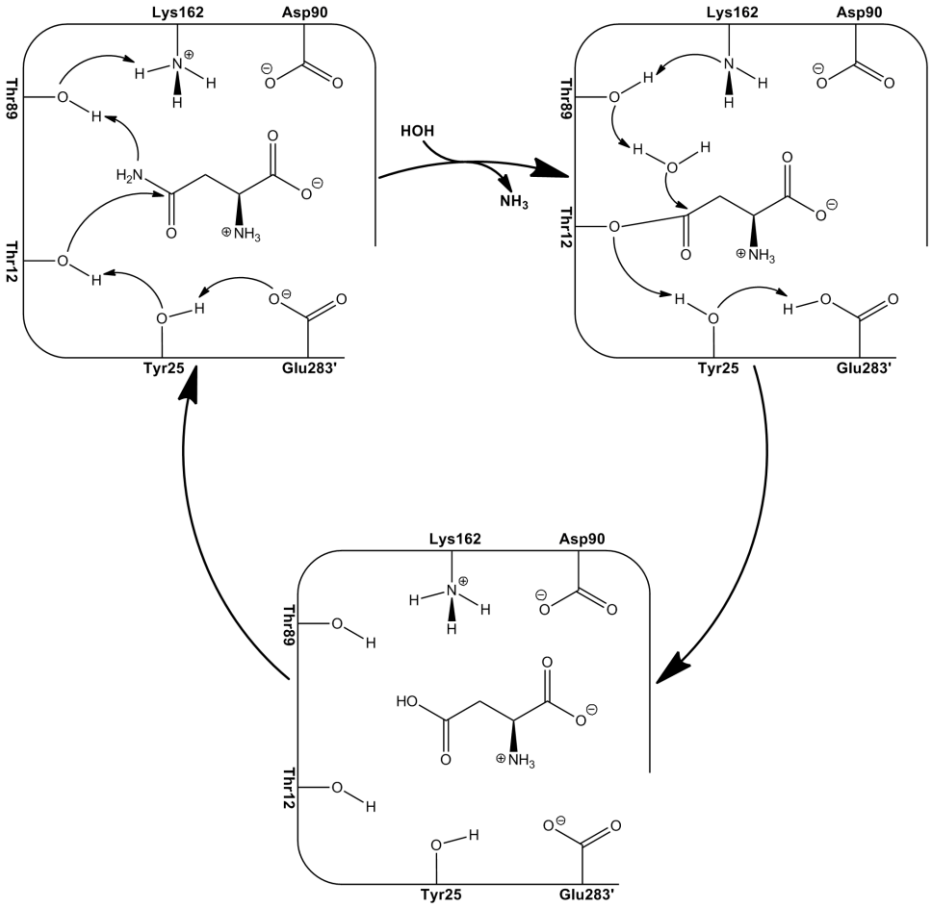


Figure 33. Currently proposed catalytic mechanism of L-asparaginase II.

This mechanism was proposed taking into account the current knowledge about the mechanism of serine proteases that share a similar catalytic triad (Ser-Ser-Lys, instead of a Thr-Thr-Lys). Although many authors point out that the mechanism of serine proteases and L-asparaginase should be very similar, there are many other aspects that led us into thinking that they can be very different. For instance, in the mechanism of serine proteases, it is the serine residue that binds to the substrate, while in L-asparaginase it is proposed to be a threonine. Although the reactivities of threonine and serine are very similar (pKa ~ 13), the presence of a methyl group nearby the hydroxyl group of threonine may difficult the formation of the acyl-enzyme intermediate. Secondly, even though there is an X-ray structure of L-asparaginase showing the acyl-enzyme intermediate, with Thr12 covalently bound to the substrate, this may result from an alternative reaction pathway, different from the wild-type, due to the mutation itself. In

fact, this intermediate, in which Thr12 is bound to the substrate, was exclusively obtained when Thr89 was mutated by a valine. In addition, when Thr12 is mutated by an alanine the catalytic process is drastically reduced. However, a decrease of 1000 fold in the rate constant[357] corresponds to an increase of only 4.2 kcal/mol in the rate-limiting step. Therefore, it is not strange that the mutation of a residue at the active site, involved in the fundamental h-bond network, destabilizes the TS by 4 kcal/mol. However, that does not mean that the residue is participating directly in the reaction but instead that the residue is very close to the reactive center.

All of these results raise many questions regarding the currently accepted mechanism of L-asparaginases and therefore require an urgent re-evaluation. To this purpose, we built a model of the enzyme containing all the residues of one dimer that participate directly or indirectly in the reaction. The model used to study the catalytic mechanism of L-asparaginase II from E. coli was based on the x-ray structure that is available on the protein databank with the code 3ECA.[352] The structure shows that the enzyme is a tetramer composed by four identical subunits (A, B, C and D) that interact with each other in the form of two intimate pairs of subunits. In this respect the tetramer is regarded as a dimer of identical intimate dimers (here called AC and BD), each one containing two active sites that are located at the interface between both subunits. With this model we have explored many hypotheses for the catalytic mechanism and calculated accurate activation and reaction energies. The results altogether point to a catalytic pathway that is different from the earlier proposals.

5.2. Methodology

5.2.1. Building the Model

The model used in this study was centered in one active site of the enzyme that is located between subunit A and C. The model contains the substrate (an asparagine molecule) and Gly10, Gly11, Thr12, Ile13, Ala14, Ser23, Asn24, Tyr25, Thr26, Val27, Gly57, Ser58, Gln59, Asp60, Gly88, Thr89, Asp90, Thr91, Gly113, Ala114, Met115, Arg116, Thr161, Lys162 and Thr163, all from chain A, and Gly248, Asn248, Leu249, Ala282, Glu283 and Val284, from chain C. The model also includes one water molecule that is conserved in the active site of L-asparaginase and it is located between the substrate, Thr89 and Lys162 (Figure 34).

Hydrogen atoms were added to the model using the software GaussView. Conventional protonation states for all amino acids at pH 7.0 were adopted, except for

Lys162 whose side chain was kept in the form of NH_2. This protonation state was deliberately chosen because earlier experimental results have shown that Lys162 is deprotonated at neutral pH[358]. The truncation of the model was done in the terminus of each part included in the model, i.e in the carboxylate and the amino groups, which were modeled as neutral, to avoid introducing artificial charges due to the truncation process. Also, some of the residues at the terminus of each part, which did not seem to interfere with the active site, were mutated to alanine, in order to diminish the number of atoms in the model. These residues include: Val27, Arg116, Thr161, Thr163, Leu249 and Val284. The final model spans a total of 416 atoms, including all those that interact directly with the substrate and are required to maintain the main scaffold of the surrounding region of the active site as it is observed in the X-ray structure 3ECA.

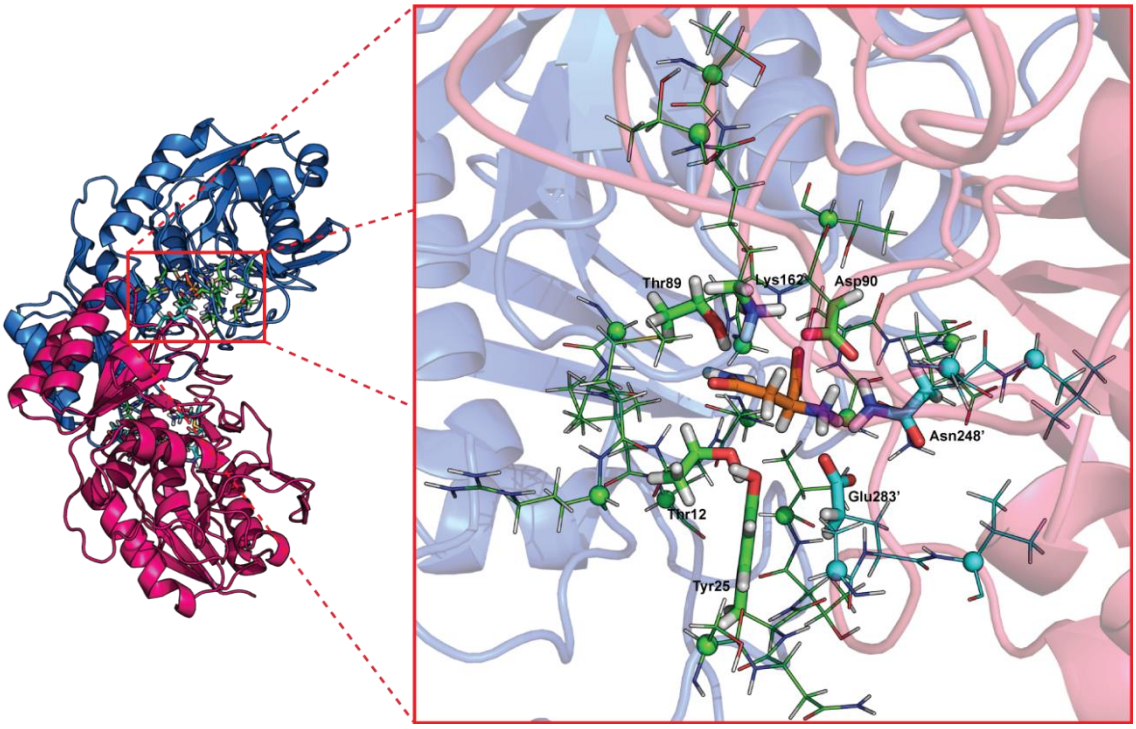


Figure 34. Left: cartoon representation of the L-Asparaginase II dimer (PDB ID: 3ECA), with both active sites shown in sticks. Right: QM/QM model. The high layer is illustrated in sticks (77 atoms) and the low layer in lines (339 atoms). The frozen atoms are depicted as spheres. Carbons are colored differently for residues that belong to different subunits: green for subunit A, cyan for subunit C and orange for the substrate.

Since the geometry optimizations can lead to a conformational reorganization of the terminal amino acids, and this may lead to localized unfolding in the model, we have chosen to freeze some atoms in the model in order to maintain it close to the X-ray structure. The C_α atoms that were frozen are represented in Figure 34 as spheres.

5.2.2. Theoretical Methods

Given that the model system is very large (over 400 atoms) and geometry optimizations are very time-consuming, we have resorted to the ONIOM methodology to perform geometry optimizations.[64, 361] This method allows the division of a system in several regions, each one studied with a different theoretical level. The accuracy of the method depends on the chosen regions, the theoretical level used in each of them and the coupling scheme between the methods. According to the ONIOM methodology, we have divided the system into two overlapping layers, designated as the high-level and low-level layers. The high-level layer includes all the atoms from the substrate (asparagine) and all the residues that are directly or indirectly involved in the reaction, i.e. Thr12, Tyr25, Thr89, Asp90 and Lys162 from chain A and Asn248 and Glu283 from chain C. Details about the exact atoms included (beyond the whole side chains) can be seen in Figure 34. This layer accounts for a total of 77 atoms. The low-level layer contains all the remaining atoms of the model and has a total of 339 atoms. The atoms described at the low-level of theory do not undergo significant geometry changes during the studied reactions, but it is still important to take them into account to get the correct orientation of the active site residues and to include the medium/long range interactions between the enzyme and the substrate. We used hydrogen atoms as link atoms to complete the valences of the bonds spanning between the two layers.

The geometry of the high-level layer was optimized with the higher theoretical level (DFT). The B3LYP functional was chosen, since it is known to give very good results for organic molecules.[362-365] The 6-31G(d) basis set was employed, as implemented in Gaussian 09.[366] The inclusion of diffuse functions in the basis set for geometry optimizations was investigated before.[367] The conclusion was that the corrections to the geometry were very small, and corrections in energy differences (such as energy barriers or energies of reaction) were negligible upon the calculation of single point energies with a more complete basis set. Therefore, it seems inadequate from a computational point of view to include diffuse functions in geometry optimizations, considering the inherent increase in computing time that they would cause. The low-level layer was treated with the semi-empirical method AM1.[368]

The model, with the aforementioned constrains, was fully optimized with the Gaussian 09 standard parameters. Subsequently, several hypotheses for the reaction mechanism were explored through linear transit scans along the reaction coordinates implicated in each studied reaction. The transition states were subsequently fully geometry-optimized, starting from the structure of the higher energy point of the scans.

The reactants and the products, associated with it, were determined through internal reaction coordinate (IRC) calculations. The transition state structures were all verified by vibrational frequency calculations, having exactly one imaginary frequency with the correct transition vector, even using frozen atoms, which shows that the frozen atoms were almost free from steric strain.

The energies of the minima and transition states were additionally calculated at the M06-2X/6-311++G(2d,2p) level in the high layer and M06-2X/6-31G(d) level in the low layer.[369] These single point calculations also accounted for the long-range contribution of the remaining enzyme through the inclusion of dielectric continuum (IEF-PCM), as implemented in Gaussian 09.[370] This feature is of particular importance to the study of enzymatic catalysis because the use of a continuum model is normally taken as an approximation to the effect of the long-range global enzyme environment in a reaction. A dielectric constant of ε=4 was chosen to describe the protein environment of the active site in agreement with previous suggestions.[371-372] Anyway, the effect of the continuum is quite insensitive to the precise choice of the value for the dielectric constant.

The atomic charges distributions were calculated at the B3LYP level employing a Mulliken population analysis scheme, using the 6-31G(d) basis set.

5.3. Results and Discussion

In order to study the catalytic mechanism of the L-asparaginases by computational means, we based our study on the X-ray structure that is available in the protein databank with the code 3ECA.[352] In the active site of this structure we have the final product of the reaction, and we believe that such configuration should not be very different from the one of the reactants. Therefore, we substituted the carboxylic group of the side chain of the acidic glutamate (product of the reaction) by an amide group in order to obtain the substrate, asparagine (the reactant of the reaction). The full enzyme structure was then minimized by molecular mechanics in order to improve the geometry of the substrate in the active site and to remove any steric hindrance that may occur after the addition of the hydrogen atoms.

The optimized geometry revealed that the substrate is wrapped in a net of hydrogen bonds provided by several residues of the active site (Figure 35). In such state, the amide group of the substrate becomes constrained between Thr89 (1.93 Å) and Thr12 (3.08 Å), forcing it to acquire a planar shape, perpendicular to the side-chains of those amino acids. Such configuration is reinforced by the presence of a conserved

water molecule (WAT) that interacts very closely with the carbonyl of the amide group (2.15 Å). The position adopted by WAT and Thr89 is very important to this end and it is ensured by the conserved Lys162 that interacts very closely with both of them by two hydrogen bonds (Lys162-Thr89: 2.22 Å, Lys162- WAT: 1.87 Å). The central role of Lys162 in these interactions is only possible due to the neighbor Asp90 (2.06 Å) that behaves almost as an anchor to Lys162 and maintains its position in the active site. Asp90 also interacts with the amino group of the substrate (1.64 Å), and with Asn248 (1.96 Å) reinforcing the net of hydrogen bonds in the active site.

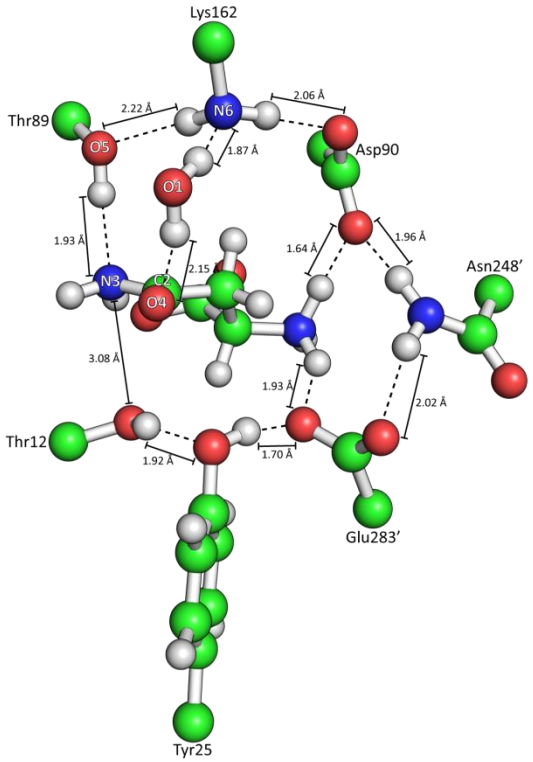


Figure 35. Optimized structure of the reactants of the reaction. (For clarity, only some of the high level atoms are shown.)

Thr12 lies in the bottom part of the active site and, together with Thr89 and WAT, interacts very closely with the amide group of the substrate. However, contrarily to what happens with Thr89, Thr12 does not establish a hydrogen bond with the atoms of the amide group. Instead the oxygen from the hydroxyl group points towards carbon C2, which has been proposed to be involved in the formation of the acyl-enzyme intermediate (3.08 Å). This configuration is favored by the proximity of the conserved Tyr25 that orients the hydroxyl group in its direction (1.92 Å) and by Glu283 that forces Tyr25 to occupy such position (1.70 Å). In the same way to what happens with Asp90, Glu283

also interacts with the amino group of the substrate (1.93 Å) and with Asn248 (2.02 Å). These interactions seem to be important for the correct alignment of all the residues of the active site and, at the same time, to force the substrate to acquire a specific conformation that might be required to start the reaction.

In our first attempts to study the catalytic mechanism of L-asparaginase, we tried to follow all the possible pathways that could lead to the formation of the acyl-enzyme intermediate involving Thr12 as it is found in the mutated X-ray structure 4ECA.[358] We have located such intermediate 11 kcal/mol above the initial reactants of the wild type enzyme. The overall barrier to reach the acyl-enzyme intermediate was 25 kcal/mol in the wild-type enzyme, obtained using B3LYP/6-31G(d):AM1. However, any attempt to progress further from the intermediate always involved barriers over 50 kcal/mol. This means that this pathway is definitely not viable in the wild type enzyme. In any case it is important to note that these results do not contradict the experimental findings. Contrarily, they give us a plausible explanation for the experimental trapping of the covalently-bound intermediate in the Thr89Val mutant structure because they show us that it is possible to form the covalent intermediate (assuming that its formation would have a comparable activation energy in the mutant and a smaller reaction energy) but that it is not possible to progress further from the covalent intermediate towards the final products.

Facing these results, we move forward in exploring other pathways that could be more feasible from the energetic point of view. In this process, we found that the most favorable pathway involves the nucleophilic attack of the substrate by a water molecule, and this occurs prior to the release of ammonia, contrarily to what has been proposed so far in the literature.

5.3.1. Step 1 − Nucleophilic Attack of the Water Molecule

The water molecule that attacks the substrate in our proposal is conserved in many X-ray structures that are available in the protein databank.[352] In those structures, this molecule is trapped between Lys162, Asp90 and the substrate. In the optimized geometry of the complex that was built for this study, the water molecule is deviated by about 2 Å from the one seen in the X-ray structure 3ECA. This position was chosen based on a 5 ns molecular dynamics simulation, where we have seen that, after transforming the product present in the crystal into the substrate, the water molecule moves to a position close to the one shown in Figure 35, and moves further to that position upon geometry optimization. In this new position it makes a hydrogen bond with

Lys162 (1.87 Å) and is particularly close to carbon C2 of the substrate that has been proposed to be involved in the formation of the acyl-enzyme intermediate (3.06 Å) (Figure 36). Such interaction is favored by the position of the amide group of the substrate that is stabilized by the network of hydrogen bonds that are available on the active site and provides a close contact, almost free of steric hindrance.

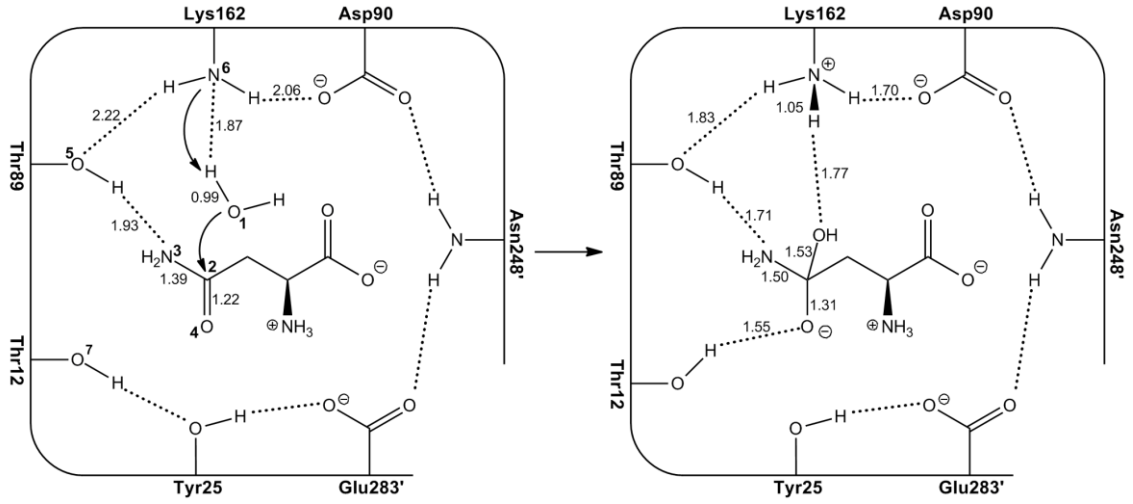


Figure 36. First step of the catalytic mechanism of L-asparaginase II.

As the reaction moves from the reactants to the products, carbon C2 changes the hybridization from sp2 to sp3, as the water molecule approaches it. Simultaneously, Lys162 becomes positively charged (-0.01 a.u. in the reactants and 0.70 a.u. in the products, for the $–NH_3$ group). At the transition state, the OH bond of the water molecule is elongated to 1.31 Å, the proton-nitrogen distance is 1.22 Å and the water oxygen-C2 distance is 1.97 Å. Additionally, the bond between oxygen O4 and carbon C2 elongates (1.25 Å), as it changes from a double to a single bond and the oxygen becomes negatively charged. Thr12 starts moving away from Tyr25 towards this oxygen, in order to stabilize this charge (Figure 37). The transition state of this reaction is characterized by one imaginary frequency at 749 icm^{-1} and reveals that the proton transfer from the water to Lys162 occurs simultaneously with the nucleophilic attack of the water molecule to carbon C2 of the substrate.

At the end of this reaction, carbon C2 adopts a tetrahedral configuration and the hydroxyl group of the water molecule becomes covalently bound to it (Figure 36). The double bond between C2 and O4 changes to a partial single bond (distance changes from 1.22 Å in the reactants to 1.31 Å in the products) and oxygen O4 becomes negatively charged (-0.72 a.u.). This charge is promptly stabilized by Thr12, which makes

a hydrogen bond with it (1.55 Å). In the course of this reaction, the distance between carbon C2 and the amino group increases from 1.39 Å to 1.50 Å and the hydrogen bond between the latter and Thr89 decreases from 1.93 Å to 1.71 Å. The same pattern is also observed with the hydrogen bond between Thr89 and Lys162, which decreases from 2.22 Å to 1.83 Å due to the strengthening of the Lys162-Thr89 H bond that changes from dipolar to ionic. All of these patterns clearly indicate that the active site is ready to catalyze the proton transfer from the positively charged Lys162 to the amino group that is attached to carbon C2 of the substrate in the next step.

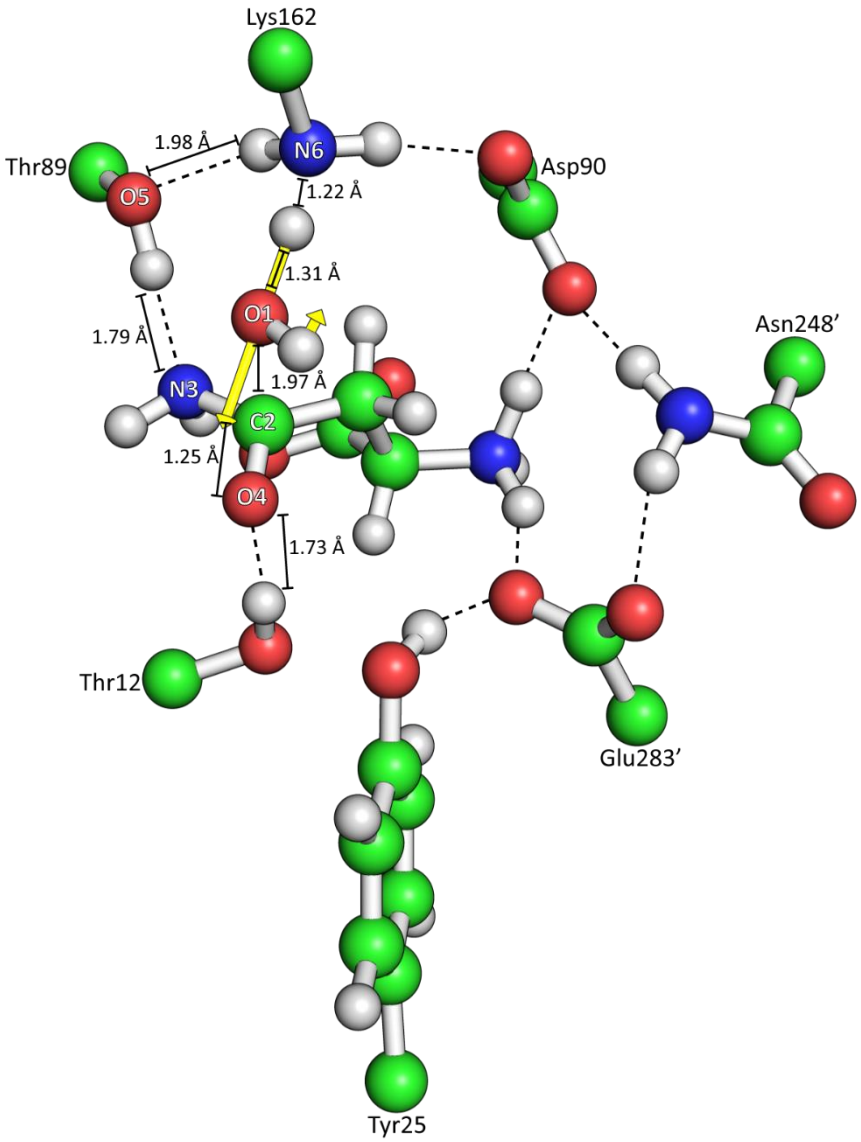


Figure 37. Transition state from the first step of the reaction mechanism of L-asparaginase II. The main vectors are represented with yellow arrows (749.3018i). (For clarity, only some of the high level atoms are shown.)

It should be emphasized that the role of Lys162 in step 1 is very important. Apart from being directly involved in the reaction, it guarantees the correct orientation and alignment of the water molecule as well as that of Thr89, which will be required for the next step. In addition, at the end of this reaction Lys162 becomes positively charged and interacts very closely with the anionic Asp90 (1.70 Å).

The reaction involved in step 1 has an activation energy of 20.2 kcal/mol and the reaction is endothermic (11.0 kcal/mol). In spite of high, the activation energy is still acceptable when compared with the experimental value regarding the kinetics of L-asparaginases (around 16 kcal/mol). The relatively high activation energy of this reaction may be explained, at least partially, by sub-optimal position of the water molecule to perform the nucleophilic attack, as well as by the position of Thr12 which must break its hydrogen bond with Tyr25 in order to stabilize the nascent oxoanion in the substrate. The inclusion of a water molecule between Tyr25 and Thr12 might lower the barrier significantly. We have seen that such water molecule makes a hydrogen bond with the carbonyl oxygen of the substrate already in the reactants. However, we decided not to include it in the final calculations because we are unsure if that water should be present in physical conditions. The position, and particularly, the orientation of the water molecule is very important for the success of the reaction, as well as for the stabilization provided by Thr12 to oxygen O4. Other factors such as long-range interactions from the missing part of the enzyme or conformational/dynamic effects, as well as the limited accuracy of DFT, may be responsible for the small difference of 4 kcal/mol between theory and experiments.

5.3.2. Step 2 – Formation of Ammonia

As mentioned before, all the active site residues are pre-organized to facilitate the formation of ammonia. Thr89 makes two hydrogen bonds, one with the amino group of the substrate (1.71 Å) and another with Lys162 (1.83 Å), favoring in this way the concomitant transfer of both protons (from Thr89 to the amide group and from Lys162 to Thr89) that ultimately leads to the protonation and release of the amino group attached to carbon C2 of the substrate (Figure 38). Thr12 performs a hydrogen bond with oxygen O4 and Lys162 interacts in the same way with oxygen O1, which belonged to the water molecule and is now covalently bound to carbon C2.

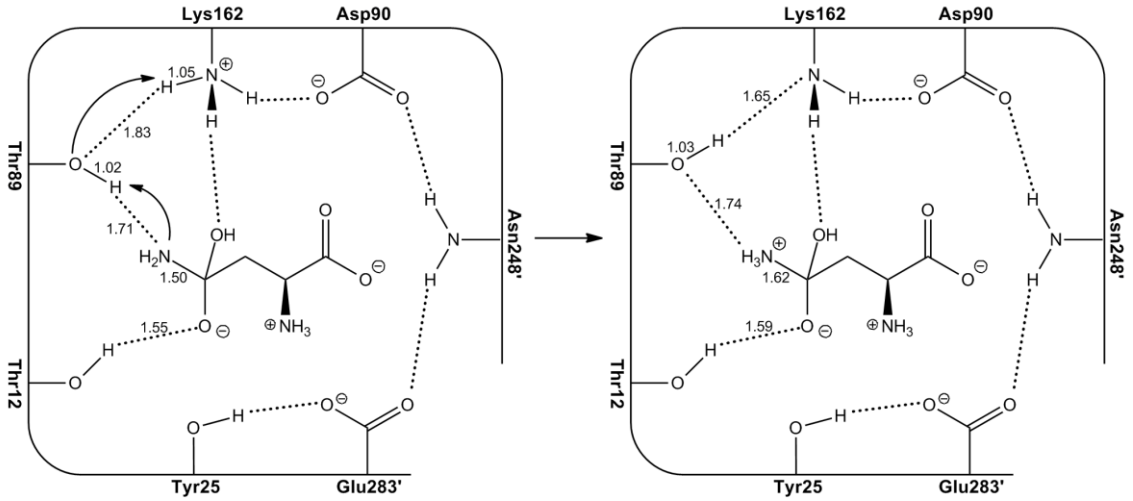


Figure 38. Second step of the catalytic mechanism of L-asparaginase.

As the reaction proceeds, the hydrogen from Thr89 moves to the nitrogen N3 of the amide group of the substrate in order to form the ammonia molecule. At the same time, the hydrogen from Lys162 draws closer to the oxygen from Thr89. In the transition state (Figure 39), it is possible to observe that the proton of Lys162 is half-way through to Thr89 (the distance from nitrogen N6 to the hydrogen is 1.21 Å while that of the hydrogen to the oxygen O5 is 1.31 Å). In the same way, the distance between the hydrogen from Thr89 and O5 has increased from 1.02 Å in the reactants to 1.56 Å in the transition state, while the distance between this hydrogen to nitrogen N3 has decreased to 1.08 Å. The transition state of this double proton transfer is characterized by an imaginary frequency at 515 icm-1, and reveals that both protons are transferred simultaneously with a relatively low activation barrier (3.5 kcal/mol), as it was proposed before.

In the product of the reaction, one ammonia molecule is obtained but it does not unbind from the substrate, as it was expected. Instead, this group remains very weakly bound to carbon C2 of the substrate (1.62 Å). This may explain why the reaction is endothermic by 4.9 kcal/mol. The total charge of the NH_3 group is 0.33 a.u. and not zero because it is still bound to the substrate. One cause for such behavior may be related to the ionic bond between oxygen O4 and Thr12 (1.59 Å) that helps stabilize the tetrahedral intermediate centered on carbon C2, and will be transformed into a weaker dipolar hydrogen bond upon release of ammonia. It is also observable that the proton from Lys162 is now bound to Thr89 (distance between N6 and the proton in the products is 1.65 Å and the distance between the same proton and O5 is 1.03 Å).

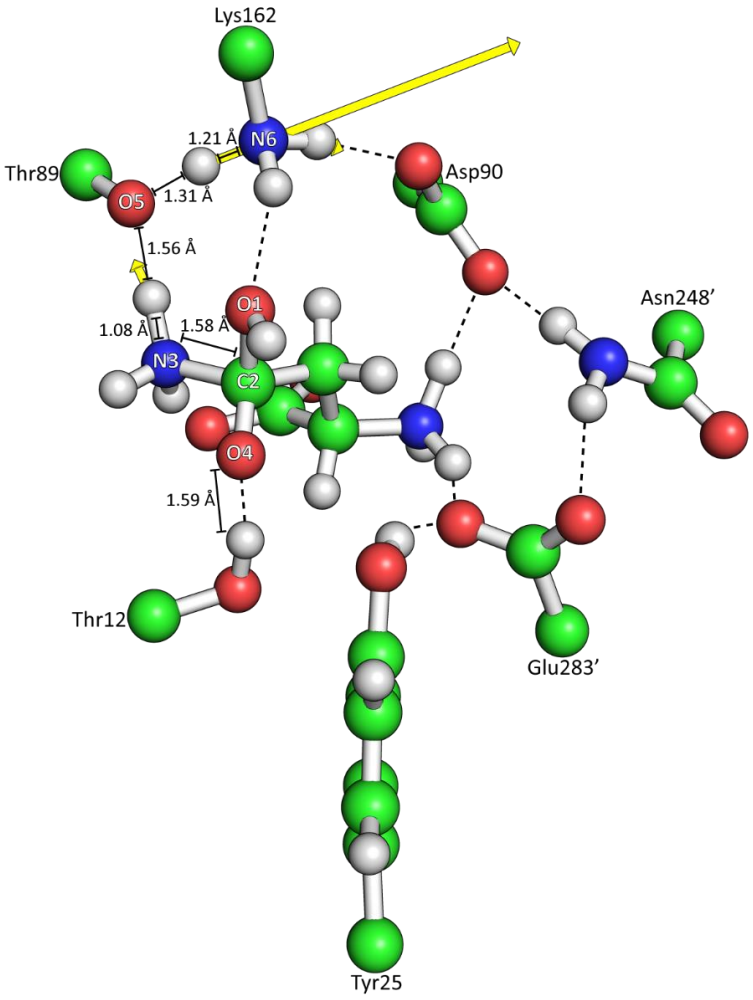


Figure 39. Transition state from the second step of the reaction mechanism of L-asparaginase II. The main vectors are represented with yellow arrows (515.8669i). (For clarity, only some of the high level atoms are shown.)

It is also interesting to note that the structure of the reactants, transition state and products can be almost superimposed with a root means square deviation of 0.3 Å, considering only the QM region. This means that the active-site is extremely well pre-organized to move from the reactants to the transition state and to the products, with minimal reorganization energetic cost. In this regard, the positions occupied by Thr89 and Lys162 are very important but also that of Thr12, which continues to stabilize the negatively charged oxygen O4 (-0.69 a.u.), and consequently the tetrahedral intermediate.

5.3.3. Step 3 – Enzymatic Turnover

The last step of the reaction involves the formation of the product of the reaction (glutamate) and the release of one ammonium molecule (Figure 40).

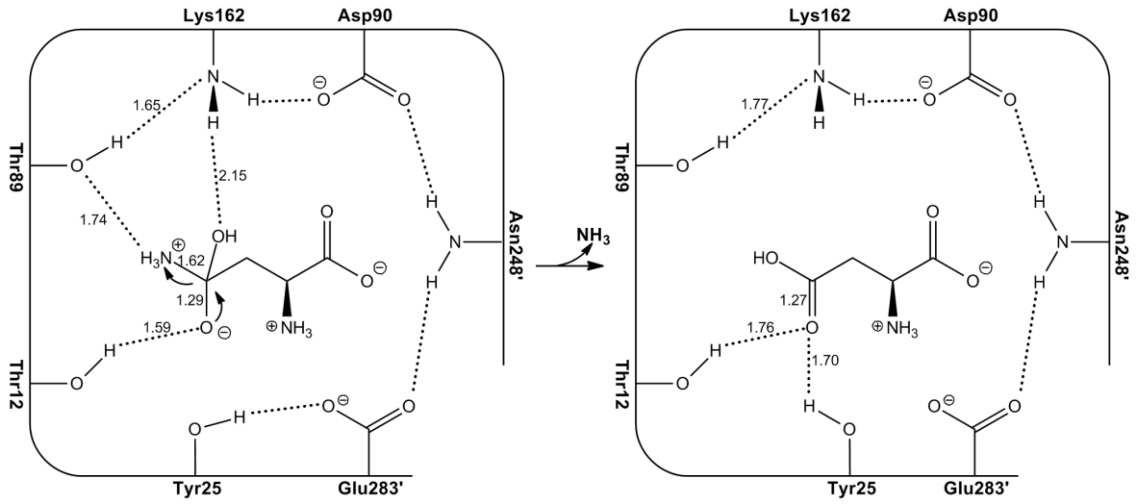


Figure 40. Third step of the catalytic mechanism of L-asparaginase.

In the optimized geometry of the reactants, it is clear that the interaction between the NH_3 group and carbon C2 of the substrate remains very stable (1.62 Å). The NH_3 and the hydroxyl groups of the substrate side chain interact very closely with two active site residues, Thr89 (1.74 Å) and Lys162 (2.15 Å) respectively, through two hydrogen bonds. In addition, these two residues interact with each other very closely (1.65 Å) through a hydrogen bond, creating a net of hydrogen bonds through which the positive charge of the NH_3 group of the intermediate can spread. Thr12 continues to interact very closely with oxygen O4 of the substrate (1.59 Å) supporting the stability of the tetrahedral intermediate.

In order to trigger the release of ammonia, we scanned the distance between the NH_3 group and carbon C2 of the substrate. The resulting dissociation of an ammonia molecule is very favorable and only has an activation energy of 3.8 kcal/mol. At the transition state it is possible to see that there are two changes occurring simultaneously (Figure 41). As the NH_3 group dissociates from carbon C2 (distance changes from 1.62 Å to 1.92 Å) forming the aspartate molecule, the distance between oxygen O4 and carbon C2 decreases (1.25 Å), implying the formation of a double bond between both atoms. The hydrogen bonds between Thr89 and Lys162 and between O4 and Thr12 are preserved during the reaction. The transition state structure of this reaction is characterized by an imaginary frequency of 179 icm^{-1}.

At the end of this reaction, the aspartate molecule is generated and the ammonia molecule dissociates from the intermediate. After optimization of the products, the NH_3 molecule is found 4.86 Å away from carbon C2 of the substrate. Carbon C2 and oxygen O4 of the substrate are now connected through a double bond (1.27 Å). The hydroxyl

group of Thr12 continues to make a hydrogen bond with oxygen O4 (1.76 Å) as well as Tyr25 (1.70 Å), which turned away from Glu283 in order to interact with it. The full reaction is rather exothermic (-25 kcal/mol), which favors the formation of glutamate.

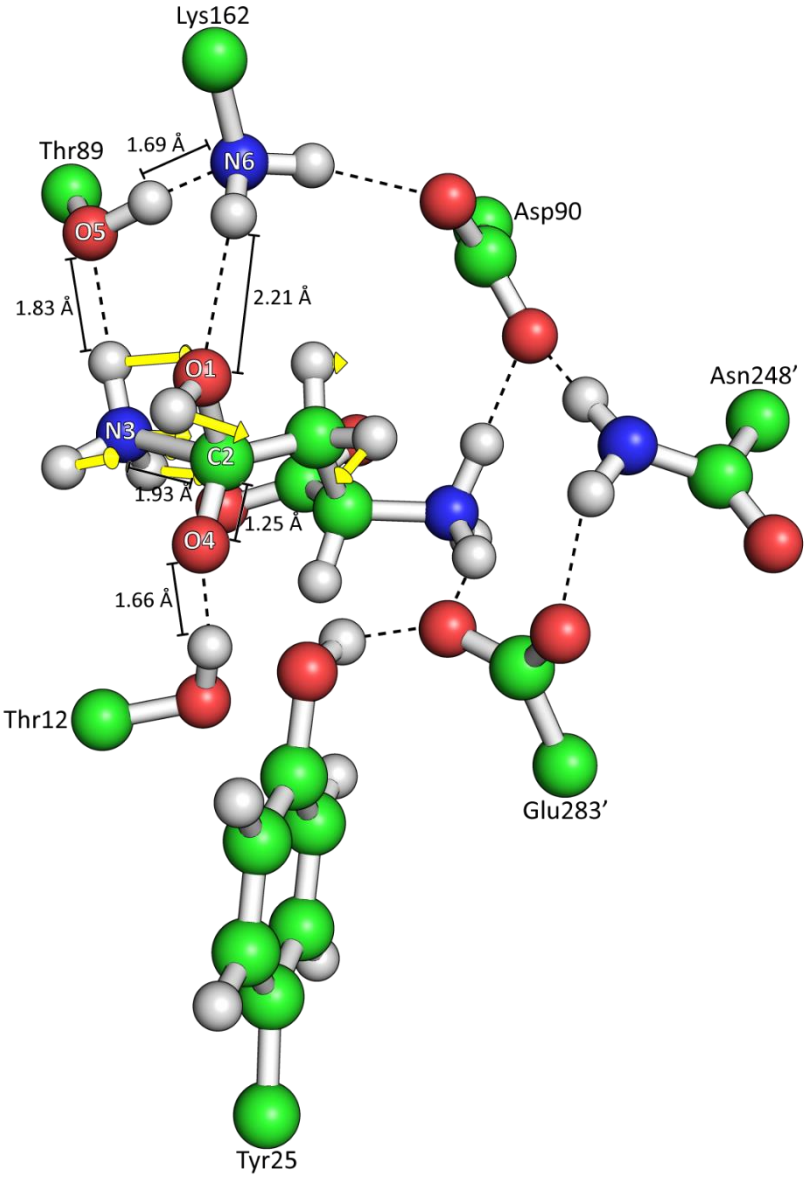


Figure 41. Transition state from the Third and last step of the reaction mechanism of L-asparaginase II. The main vectors are represented with yellow arrows (-179.4042i). (For clarity, only some of the high level atoms are shown.)

In the end of these three steps, the reaction is completed and asparaginase can release the aspartate molecule from the active site to the solvent. Subsequently, the protein is ready to take another molecule of asparaginase and start the reaction all over.

5.4. Conclusion

The computational results addressed to the study of the catalytic mechanism of L-asparaginase II have shown that the full mechanism involves three sequential steps and requires the nucleophilic attack of a water molecule to the substrate prior to the release of ammonia (Figure 42). The first step of the mechanism involves the formation of a tetrahedral intermediate that results from the nucleophilic attack of the water molecule to carbon C2 of the substrate. In the course of this reaction, Lys162 receives the proton from the water molecule and becomes positively charged. The second step of the reaction involves a concerted double proton transfer from Lys162 to Thr89 and from Thr89 to the substrate, which results in the formation of ammonia. However, the ammonia molecule is still weakly bound to the substrate at this stage. This reaction requires the direct participation of Thr89, which serves as an intermediate in the proton transfer between Lys162 and the substrate. Such configuration is believed to be a result of the stabilization of the tetrahedral intermediate, in which Thr12 and Tyr25 play an active role. The last step of the reaction is very straightforward and involves the dissociation of the ammonia molecule. This reaction is very exothermic, and once it is complete and the product of the reaction leaves the active site, the enzyme is ready for the next turnover. These results also show that the full reaction is almost thermoneutral, a condition that is in agreement with the available experimental results, which show that this enzyme is capable of catalyzing the reaction in both directions (Figure 43).

Another conclusion that the results provide regards the importance of the extensive network of hydrogen bonds between the residues of the active site and the substrate. Two specific regions can be highlighted. One region interacts with the top of the substrate, and involves the catalytic triad Thr89, Lys162 and Glu90. These residues were shown to be important for the catalytic activity of the enzyme, but also to orient the water molecule in relation to the substrate. The second group of residues involves Thr12, Tyr25 and Glu283. These residues are not directly involved in the reaction, but are very important for the stabilization of the tetrahedral intermediate of the substrate that is essential for the reaction. This might explain why the mutation of Thr12 to an Ala makes the enzyme almost inactive, but if it mutated by a Ser residue the normal pathway is maintained[373]. Although these two triads of active site residues are often regarded as independent regions, they are in fact associated with each other by two hydrogen bonds promoted by Asn248. Since both Thr12 and Tyr25 are included in the mobile loop, which allows the substrate to enter and leave the active site, we can conclude that the presence of these residues is also very important since they allow the enzyme to enclosure and hold tight the substrate in the binding pocket, by approaching the previous indicated

regions close together[374]. The results obtained with this study show that the formation of the acyl-intermediate with Thr12 is not energetically favored, and therefore should not be present in the natural pathway of the wild-type enzyme. However, there has been a great debate about this intermediate in the literature. In many instances, it has been proposed to be a crucial intermediate in the catalytic mechanism of this enzyme. We must stress that this acyl-enzyme intermediate was obtained only when Thr89 was mutated by a valine. In this study we have shown that Thr89 is an active player on the reaction (Figure 42) and it is fundamental for the network of hydrogen bonds that forces the substrate to acquire a productive conformation, as well as for the correct orientation of the water molecule that is essential for the reaction. This means that, when Thr89 is mutated by a valine, the network of hydrogen bonds is completely disrupted, and the proton transfer from the water molecule to the amino group of the substrate is precluded. Consequently, the formation of the tetrahedral intermediate is prevented and the reaction moves toward a different direction and ends up with the formation of the acyl-intermediate with Thr12. Comparing the X-ray structures 3ECA (wild type enzyme) and 4ECA (Thr89Val mutant), we can see this trend and, in particular, the different position and conformation that is adopted by Lys162, which moves outwards from the active site. Together with Thr89, Lys162 is another active player in this reaction and, therefore, when Thr89 is mutated by a valine, the catalytic triad formed by Thr89, Lys162 and Asp90 is disrupted, another condition that changes the normal pathway of the reaction. These facts altogether show that the formation of the acyl-intermediate with Thr12 may be a cause of the mutation rather than a true intermediate in the wild-type mechanism.

Another piece of evidence that supports this concept is the fact that, when Thr12 is mutated by an alanine, the enzyme becomes less efficient, but remains active (the K_{cat} of the Thr12Ala mutant is approximately 1000 times less than that of the wild-type). Even though one may think that a 1000 fold decrease in the rate constant will render the enzyme completely inactive, in reality this translates in an increase of only 4.2 kcal/mol in the rate-limiting step. It is not strange that the mutation of a residue at the active site, involved in the fundamental H-bond network, destabilizes the TS by 4 kcal/mol and it does not necessarily means that the residue is participating directly in the reaction but instead that the residue is very close to the reactive center. This suggests that this residue is not an active player on the reaction, as it has been proposed and as it was corroborated by our results.[356] In fact, Thr12, together with Tyr25, is very important for the stabilization of the tetrahedral intermediate that allows the binding of the water molecule to the substrate. If the function of one of these residues is disturbed, as with the mutation of Thr12 by an alanine, it will slow down the formation of the tetrahedral

intermediate. The last condition is perhaps what is observed when Thr12 is mutated by an alanine. Our conclusions are also supported by the fact that when Thr12 is mutated by a serine, the catalytic activity is not changed. This is completely in line with the catalytic mechanism that we propose. As the role of Thr12 is to stabilize the tetrahedral intermediate and position the substrate through H-bonding with its side-chain hydroxyl, it becomes obvious that a serine should be able to fulfill the same role with its equivalent hydroxyl group. This is the reason why the catalytic activity is undisturbed in the Thr12Ser mutant.

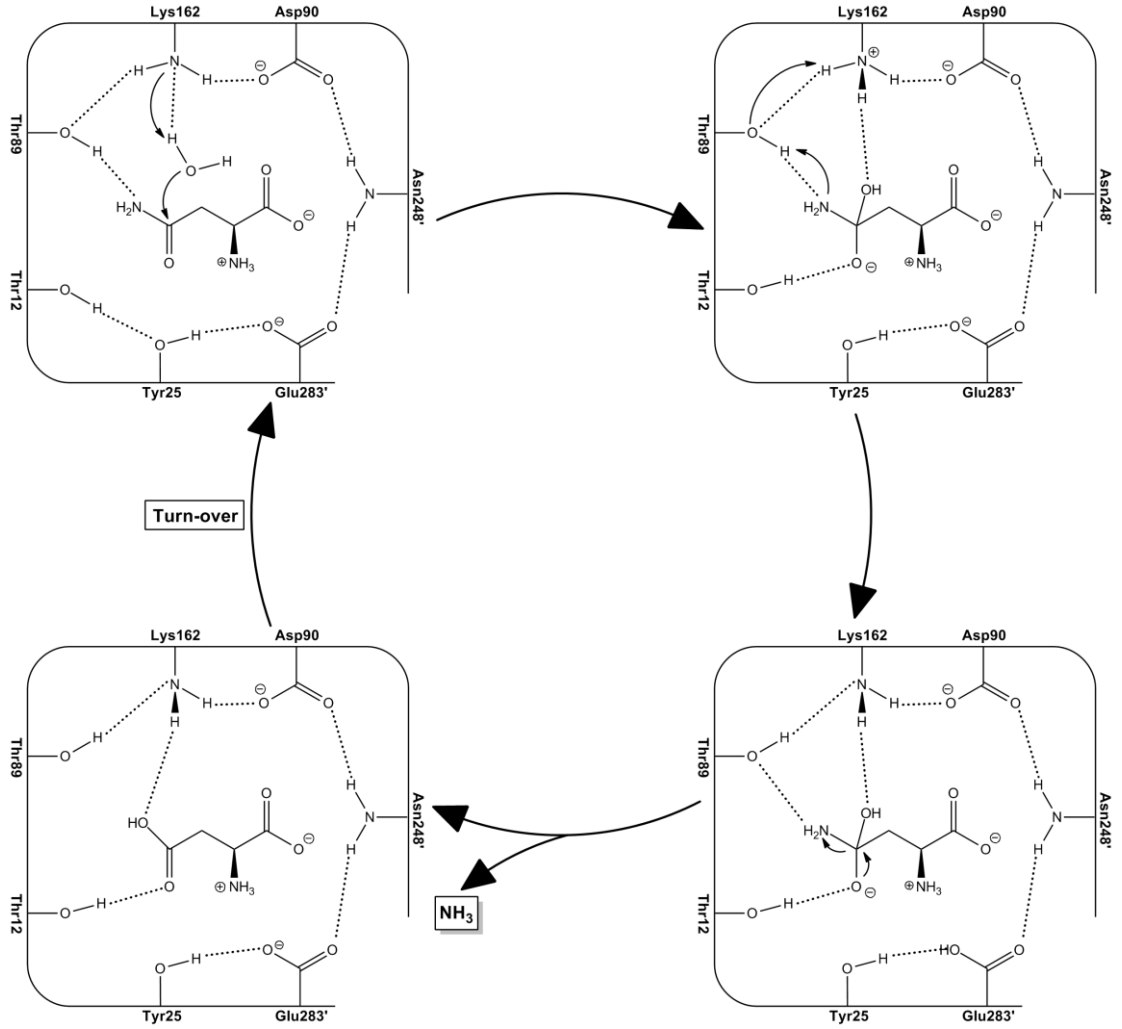


Figure 42. New proposal for the catalytic mechanism of L-asparaginase II.

We believe that the new mechanistic portrait of the catalytic mechanism of L-asparaginase II provides a new, but at the same time complementary, understanding of the catalytic mechanism of L-asparaginases that fully respects all experimental data on this subject. This knowledge is not only important in order to gather an atomistic portrait

about the catalytic activity of this enzyme, but also to acquire additional information about the transition state structures that are useful to develop new compounds capable of controlling its activity, which is of extreme importance in cancer therapy. As a chemotherapeutic drug, this enzyme can also be harmful to the normal cells in the body if it is active in the blood stream for a large period of time. For this reason, L-asparaginase therapy is accompanied by the administration of an inhibitor some time after the uptake of the enzyme. With the results we have obtained, new and improved L-asparaginase II inhibitors can be studied.

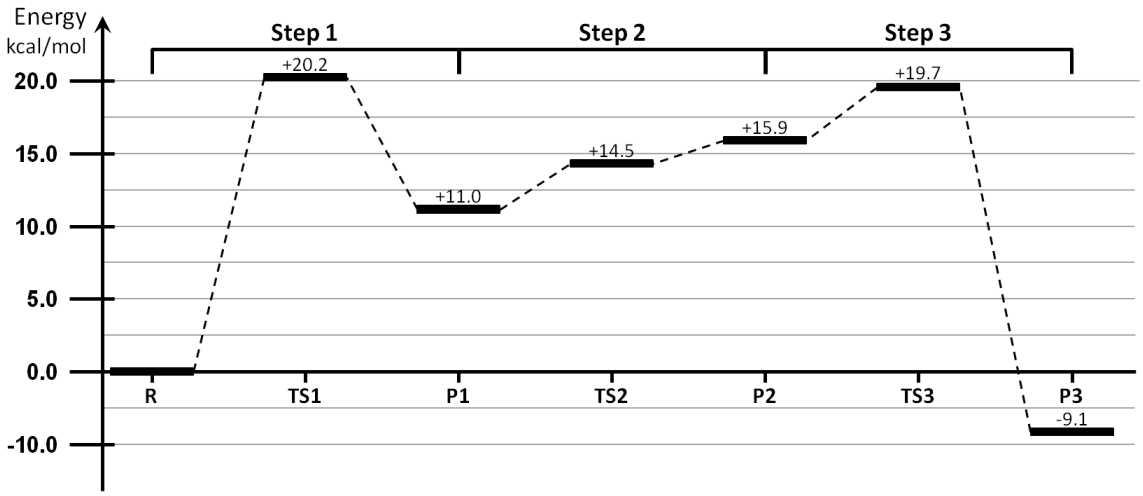


Figure 43. Energetic profile of the catalytic mechanism of L-asparaginase II.

We believe that the new mechanistic portrait of the catalytic mechanism of L-asparaginase II provides a new, but at the same time complementary, understanding of the catalytic mechanism of L-asparaginases that fully respects all experimental data on this subject. This knowledge is not only important in order to gather an atomistic portrait about the catalytic activity of this enzyme, but also to acquire additional information about the transition state structures that are useful to develop new compounds capable of controlling its activity, which is of extreme importance in cancer therapy. As a chemotherapeutic drug, this enzyme can also be harmful to the normal cells in the body if it is active in the blood stream for a large period of time. For this reason, L-asparaginase therapy is accompanied by the administration of an inhibitor some time after the uptake of the enzyme. With the results we have obtained, new and improved L-asparaginase II inhibitors can be studied.

This knowledge also provides important clues to understand the source of the catalytic activity of the enzyme, which can also be used to enhance its efficiency or modify the type of substrates, or even the reactions that can be catalyzed by these enzymes.

CHAPTER 6

DISCOVERY OF NEW DRUGGABLE SITES IN THE ANTI-CHOLESTEROL TARGET HMG-COA REDUCTASE BY COMPUTATIONAL ALANINE SCANNING MUTAGENESIS

The enzyme 3-hydroxy-3-methyl-glutaryl-CoA reductase (HMG-CoA-R) is the fundamental target for the treatment of hypercholesterolemia nowadays. The HMG-CoA-R clinical active site inhibitors (statins) are among the most widespread and profitable drugs ever sold but their side effects (myopathies, sometimes severe) still limit their use, which makes the finding of alternatives to statins a field of intense research. In this line, we address here a new strategy for inhibiting the homotetrameric HMG-CoA-R. The enzyme consists in a "dimer of dimers", each dimer having two active sites. We pursue here the inhibition of enzyme oligomerization, through drug binding to the dimer interface. We have mutated computationally 232 interfacial residues by alanine and calculated the lost in binding free energy among the monomers that build up each dimer of the homotetramer. This led to the identification of the (ten) key residues for the formation of the active dimer (Glu528, Ile531, Met534, Tyr644, Glu665, Asn686, Lys692, Lys735, Met742 and Val863). The results show that these residues are located in two specific spots of the protein with a cleft shape, whose shape and size is favorable for small drug binding. It is expectable that small molecules specifically bound to these druggable pockets will have a major effect in the oligomerization of the protein or/and in active site formation. This paves the way for the discovery of new families of inhibitors of HMG-CoA-R.

Adapted from reference [375]

For this paper Diana Gesto preformed all the calculations and analyzed the results, wrote the entire preliminary draft of the manuscript which was then reviewed by all co-authors.

6.1. Introduction

Cholesterol is a molecule of most importance in the biological world, and especially important for animals, where it is a key player in several biological processes, including

membrane fluidity and the biosynthesis of hormones. However, cholesterol has currently gained a bad reputation next to the great majority of people, especially because of its association with cardiovascular diseases, the number one cause of death in the developed world[376].

In humans, cholesterol can derive from different sources, the most important of which is the de novo synthesis, which starts with the mevalonate pathway[10]. In this pathway, two molecules of acetyl-CoA are condensed forming acetoacetyl-CoA and afterwards a new molecule of acetyl-CoA is used to create HMG-CoA. HMG-CoA is then reduced by NADPH, in a reaction catalyzed by the enzyme 3-hydroxy-3-methyl-glutaryl-CoA reductase (HMG-CoA-R), in order to form mevalonate, an important intermediate in the formation of many compounds, including cholesterol. This last step is not only the committed step of the entire pathway, but it is also the rate limiting one[38].

Statins are currently the most used drugs to decrease the levels of cholesterol in the blood through the inhibition of HMG-CoA-R. The different types of statins differ in the structure apart from the HMG-like moiety that is common to all of them. These structural differences result in different binding characteristics in the HMG-CoA-R active site that confer them different inhibitory potencies[377]. In spite of the great success of statins in decreasing blood cholesterol levels, there are significant side effects associated with their use. While some of these can cause beneficial (pleiotropic) effects (plaque stabilizing, anti-inflammatory and antithrombotic effects, etc.) others can be adverse and even lethal like rhabdomyolysis and myalgia[378]. In 2001, a statin called cerivastatin was withdrawn from the market because of several cases of rhabdomyolysis associated with it[379]. Still, in spite of all the possible side effects, it did not preclude atorvastatin from being the most profitable drug in the world over the last 8 consecutive years, which demonstrates the great demand for drugs that lower the blood concentration of cholesterol.

Until now the competitive inhibition of HMG-CoA reductase by statins has proven to be the most efficient way of decreasing the biosynthesis of cholesterol and, therefore, it will continue to be the main approach to treat hypercholesterolemia in a near future. However, it is clear that new drugs targeting this enzyme, and consequently capable of lowering blood cholesterol levels, are needed. At the same time new ways of inhibiting HMG-CoA-R are also being sought and, in this context, the dimerization inhibitors are very promising.

As it is known, each of the four active sites of HMG-CoA-R is composed of residues from two identical monomers (the whole enzyme is a tetramer, organized as a dimer or

dimers). This means that unless the dimer is formed, the enzyme cannot work properly. Inhibitors that interfere with the formation of the dimer will inhibit the biosynthesis of mevalonate, and consequently, that of cholesterol. However, in order to achieve this, we would first need to discover which residues contribute the most for the monomer:monomer association. As the monomer:monomer interfaces of the protein are very large, and bioavailable drugs should be small molecules, the drugs will only preclude a small set of interactions among the protein dimers. Therefore the drugs must specifically bind to the regions of the protein that are more important for monomer:monomer binding. To find this out, we propose the use a method called computational Computational Alanine Scanning Mutagenesis (cASM)[380]. The theory behind the cASM method is that if a residue at the dimer interface is mutated by an alanine and if we calculate the difference between the binding energy before and after the mutation was introduced, we can assess whether that residue is important or not for the formation of the interface by calculating precisely the quantitative contribution of each residue for the monomer:monomer binding free energy[380-381].

In this paper we conducted a systematic application of the cASM method to a total of 232 residues (58 in each monomer) present at the interface of the four subunits of this HMG-CoA-R, and identified the ones that have a higher contribution to the formation of the dimer. The collected data provides important clues about the association determinants for the two subunits and can serve as a basis for the development of a new class of HMG-CoA-R inhibitors designed to block the association of the two subunits, and therefore the biosynthesis of mevalonate, and that of cholesterol.

6.2. Methodology

6.2.1. Model and Molecular Dynamics Simulation

The HMG-CoA reductase model used in this study was based on the tridimensional structure of the human enzyme. In the Protein Data Bank there are over 20 structures for human HMG-CoA reductase and we have chosen the one with the PDB code 1DQA (resolution 2.00 Å)[220]. Even though there were other structures with better resolution, we preferred this one because it is complexed with the natural substrate and the CoA and NADP cofactors. This guarantees that the interface between both proteins has not been distorted and maintains the main interactions and symmetry of the wild type enzyme. In addition, from all the X-ray structures available on the PDB databank this is the only structure that has a less number of missing residues of the N-terminal region of protein.

While constructing our model we removed all atoms from the HMG, CoA and NADP molecules. We also removed all solvent molecules, except structural waters. Since it is know that the protonation states of some residues may vary depending on the environment in which they are found, we conduct a preliminary study using PROPKA[382-383]. This software predicts the pKa (and therefore the probable protonation state) for each residue in the enzyme. For our model, the results obtained showed that His672, Cys688 and Lys735 would not have the standard protonation states once inside the folded enzyme, and therefore, in our final model we changed the charges of these residues so that they would match the prediction. The hydrogen atoms were added afterwards using the LEAP program which belongs to the AMBER 9.0 software package[384]. Since the total charge of the protein was +9 we also added Cl^- ions to counterbalance it. We also added a water box (TIP3P model) with sides at least 15 Å away from any protein atom, to solvate our system, which also enables us to use periodic boundary conditions.

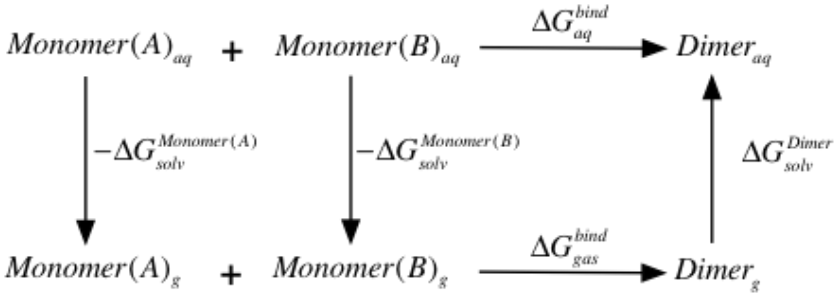
We used the FF03 force field in all calculations. Before the Molecular Dynamics (MD) simulation, we performed a three-stage energy minimization protocol using the SANDER module of AMBER 9.05, in which constraints applied to the model were progressively removed. During first stage (500 steps) the positions of all atoms of the model, except those of the water molecules, were restrained using 50 kcal mol^{-1} $Å^{-2}$ harmonic forces. For the second stage (500 steps), the same constraints were applied to all atoms except hydrogens and for the last stage (500 step) no constraints were used. Then we did a 20 ps equilibration process, in which the system was gradually warmed up until 310.0 K, using the NVT ensemble. Afterwards, a 10 ns simulation was performed on the model using NTP conditions. The SHAKE algorithm was used in order to restrain bonds involving hydrogen atoms[385]. The equations of motion were integrated at each 2 fs and a non-bonded interaction cut-off radius of 12 Å was employed with a Particle-Mesh Ewald scheme. To maintain a steady temperature of 310 K the Langevin thermostat was applied. For the posterior analysis we used only the last 4 ns of the simulation.

6.2.2. Computational Alanine Scanning Mutagenesis

The Computational Alanine Scanning Mutagenesis (cASM) combines a continuum approach to model solvent interactions with an MM-based approach to atomistically model protein–protein interactions. This methodology has been successfully applied to several complexes to predict accurately differences in binding free energies in solution between the wild-type and alanine mutated complexes ($\Delta\Delta G_{bind}$)[386-389].

$$\Delta\Delta G^{bind} = \Delta G^{bind}_{mut,aq} - \Delta G^{bind}_{wildt,aq} \quad (31)$$

The corresponding binding free energy for mutant and wild type complexes in solution (ΔG^{bind}_{aq}) were calculated using the thermodynamic cycle shown at Scheme 1 in which ΔG^{bind}_{g} is the binding free energy of the dimer in the gas , $\Delta G^{monomer(A)}_{solv}$, $\Delta G^{monomer(B)}_{solv}$ and $\Delta G^{complex}_{solv}$ are the solvation free energy of the monomer A, monomer B, and the dimer, respectively.



Scheme 1: Thermodynamic cycle used to calculate the binding free energy for mutant and wild type complexes in solution (ΔG^{bind}_{aq}) in the CompASM protocol.

The binding free energy of two molecules in a complex ($\Delta\Delta G^{bind}_{aq}$) is defined here as the difference between the free energy of the complex and those of the respective monomers.

$$\Delta G^{bind} = G^{complex} - \left(G^{monomer(A)} + G^{monomer(B)}\right) \quad (32)$$

The free energy of a dimer protein and its respective monomers ($G^{complex}$, $G^{monomer(A)}$ and $G^{monomer(B)}$)can be calculated accordingly to equation (33, summing the internal energy ($E_{internal}$), the electrostatic and the van der Waals interactions ($E_{electrostatic}$ and E_{vdW}), the free energy of polar solvation ($G_{polar\ solvation}$), the free energy of nonpolar solvation ($G_{nonpolar\ solvation}$) and, the entropic (TS) contribution for the molecule free energy.

$$G_{molecule} = E_{internal} + E_{electrostactic} + E_{vDW} + G_{polar\ solvation} + G_{nonpolar\ solvation} - TS \quad (33)$$

The first three terms in equation 33 are calculated using the Cornell force field with no cutoff. The electrostatic solvation free energy is calculated resorting to the MM-PBSA (Molecular Mechanics Poisson–Boltzmann Surface Area) approach first developed by

Massova et al.[380, 390] and that was later one adapted by Moreira et al.[391]. To this end we use the original MM-PBSA script integrated into the AMBER9 package but with different internal dielectric constants, which depend exclusively on the type of amino acid that is mutated. For the charged amino acids (aspartic acid, glutamic acid, lysine, arginine, and histidine), a constant of 4 is used. For the remaining polar residues (asparagine, glutamine, cysteine, tyrosine, serine, and threonine) not ionized at physiological pH, the internal dielectric constant should be set to 3 and for the nonpolar amino acids (valine, leucine, isoleucine, phenylalanine, methionine, and tryptophan), the internal dielectric constant is set to 2. The different internal dielectric constants account for the different degree of relaxation of the interface when different types of amino acids are mutated for alanine. This means that the stronger the interactions these amino acids establish, the more extensive the relaxation should be, and the greater the internal dielectric constant value must be to mimic these effects.

The nonpolar contribution to solvation free energy due to van der Waals interactions between the solute and the solvent and cavity formation was modeled as a term that is dependent on the solvent-accessible surface area of the molecule. It was estimated using the following empirical relation[381],

$$\Delta G_{nonpolar} = \alpha A + \beta \quad (34)$$

where, A is the solvent-accessible surface area that was estimated using the molsurf program (based on the idea primarily developed by Michael Connolly) and α and β are empirical constants for which the values of 0.00542 kcal $Å^{-2}$ mol^{-2} and 0.92 kcal mol^{-1} were used.

The entropy term (*TS*) can be obtained as the sum of translational, rotational, and vibrational components. However, it was not calculated here because it was assumed, on the basis of previous work that its contribution to $\Delta G_{aq}^{bind}$is similar for the wild type protein and for the mutants, and thus its contribution for $\Delta\Delta G^{bind}$ is neglectable[380, 392].

In the present work, the full cASM protocol was done using the compASM plug-in that was developed in our group and it is available at: (http://compbiochem.org/Software/compasm/Home.html)[393]. The full process was applied to 350 snapshots of the last 4 ns of the molecular dynamic simulation. Since the quaternary structure of HMG-CoA-R is a homotetramer compose by two active dimers, the cASM protocol was applied to both dimers interfaces.

6.2.2.1. Mutant Selection

As the dimeric interfaces of HMG-CoA-R are quite large, we restricted the residues that were tested to the ones that had the possibility of being important for binding. It was found in an earlier work[393] that any residue whose contribution for the binding energy was important (defined as contributing by more than -4 kcal/mol, or 3 orders of magnitude for the binding constant) needed to be deeply buried on the protein interface, with a buried surface area larger than 40 $Å^2$. Therefore, only the residues that had buried areas larger than 40 $Å^2$ were studied. All the others would necessarily give small contributions for the binding energy and would never be efficient drug targets. Using this cutoff we identified a total of 232 residues, 58 for each subunit (The results can be found at table 2 of the discussion section).

The mutant complexes are generated by a single truncation of the mutated side chain, replacing Cα with a hydrogen atom and setting the Cα-H direction to that of the former Cα-Cβ. The glycine residues present in the interface were not mutated, because their side-chain is smaller and their substitution would probably affect the structure tridimensional arrangement of the enzyme. Proline residues pose a similar challenge as well, and its substitution for alanine could also severely alter the structure of the protein.

6.2.2.2. SASA Analysis

The solvent accessible surface area (SASA) for all residues mutated was calculated using the Visual Molecular Dynamics (VMD) program[394] and the VolArea plug-in[395]. The standard value for the probe radius of water (1.4 Å) was used. The SASA values were calculated for the last 4 ns of the MD simulation, from a total of 2000 snapshots. We calculated the SASA for each residue considering different configurations, i.e. considering residue was free (not integrated in the protein – $SASA_{free}$), considering only the monomer in which it belongs (and disregarding the shielding effect of the other monomer of the dimer – $SASA_{monomer}$) and considering the whole dimer ($SASA_{dimer}$). The results were also expressed as a percentage of the maximum possible SASA, i.e. the SASA of the free residue.

6.3. Results and Discussion

Our purpose is to find ways of avoiding the formation of functional HMG-CoA-R, by precluding its dimerization or, at least, impairing the formation of native-fold dimmers.

As the active site is at the dimer interface, any perturbation in the quaternary structure will have severe consequences on the enzyme activity. However, it is difficult to interfere with large-molecules association using a (comparatively) very small bioavailable drug. For that purpose we need to identify the regions of the protein interface that are determinant for binding (to bind the small molecule there), and evaluate if their shape makes them are druggable.

The structure of the human HMG-CoA-R is formed by four identical monomers, which forms two dimers that coil around each other in an intricate way. The enzyme contains four active sites, two in each dimer, and they are made up of residues from both subunits. Based on this information we have analyzed by cASM both dimer interfaces of the human HMG-CoA-R that are denominated here as dimer A/B and dimer C/D (Figure 44).

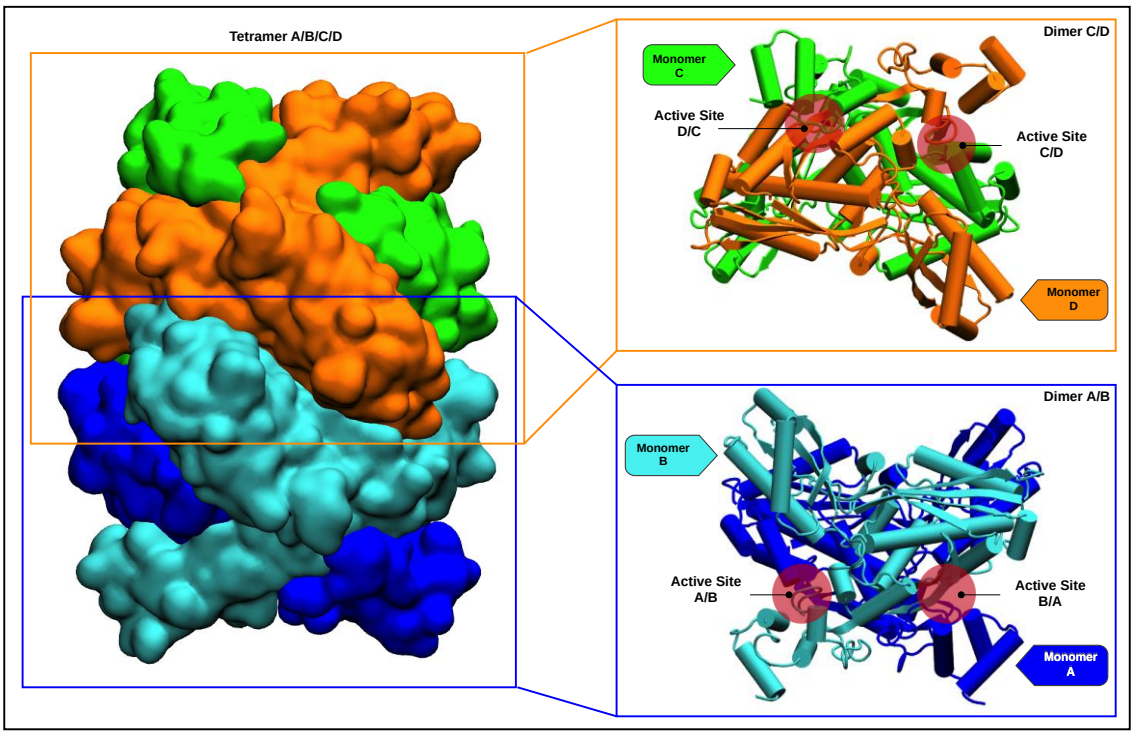


Figure 44 - Tridimensional structure for the catalytic domain of human HMG-CoA reductase (PDB code: 1DQA)

6.3.1. General Analysis of the MD Simulation

Prior to the cASM study, a general analysis of the MD trajectories generated was performed to evaluate the convergence and stability of the most relevant properties within the 10 ns simulation times considered. Properties evaluated included the potential

and energy, the temperature, the pressure, the volume, the root-mean-square deviation in the atomic positions (RMSd), etc. From these general properties, the RMSd is normally the most difficult to converge and is usually taken as a reference to assess if the MD simulation has reached equilibrium.

Figure 45 illustrates the RMSd values for the backbone Cα atoms in the MD simulations of the full tetramer and for each of the dimers. In addition, this figure also shows the RMSd values for the subset of backbone Cα atoms of the 58 interface residues that will be the subject of more detail in this study. The results show that all are well equilibrated after the initial 6 ns of simulation. The interfacial residues are perfectly stabilized. After this observation, the remaining 4 ns of simulation (from 6 to 10 ns) were taken into consideration for the ASM calculations and subsequent analysis.

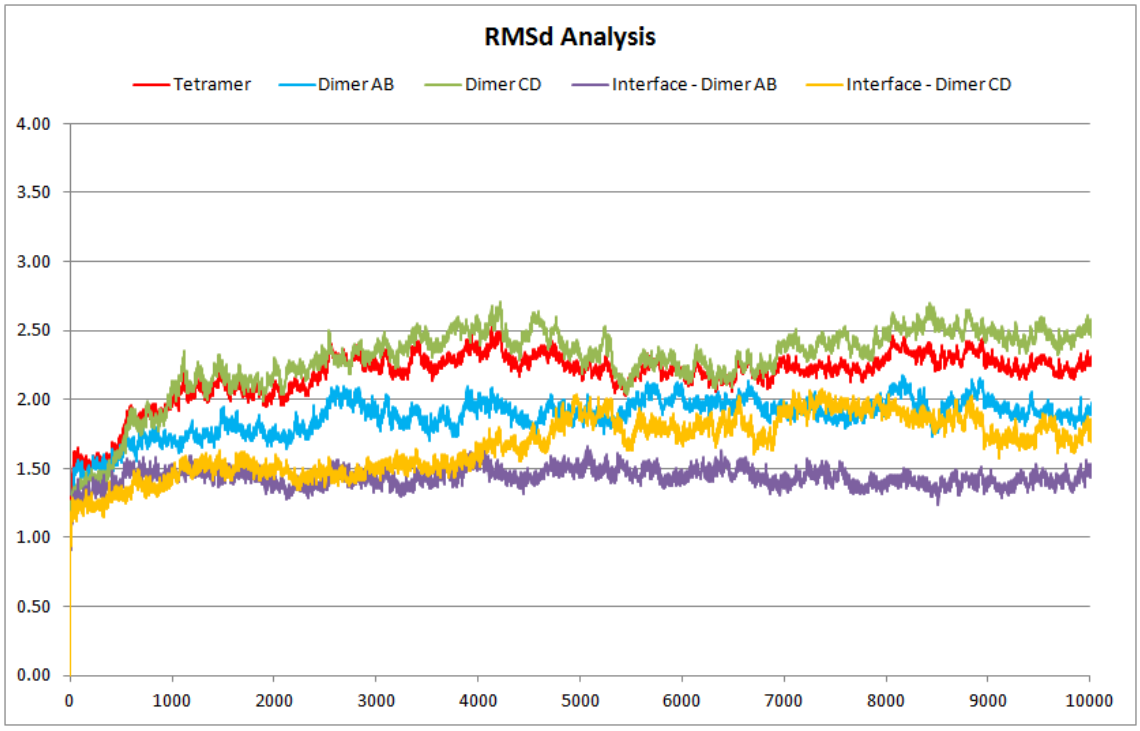


Figure 45 - RMSd analysis of the backbone Cα atoms for the tetramer, for the dimers AB and CD, and for the studied residues of the interface.

6.3.2. Analysis of the Interfaces of each Dimer of HMG-CoA-R

The interfaces of each dimer of HMG-CoA-R are quite large since all domains of the monomer participate in the interactions that join both subunits together (Figure 44). Taking this into account, we have chosen to restrict the number of residues that were

studied. This was done having in mind that the hot-spots are usually deeply buried on the protein interface and therefore only the residues whose buried surface area was larger than 40 Å² were selected and mutated. In the end of this procedure, a total of 232 residues were selected (58 of each subunit).

The energetic contribution for the dimer association was then analyzed using the cASM methodology. The calculated binding free energy difference between the wild-type and alanine mutated complexes ($\Delta\Delta G_{bind}$) is summarized at Table 3. These values correspond to the binding affinity lost when a side chain is changed for an alanine, and can be considered as the contribution of the side chain for binding (note that the "energy zero" here is the binding provided by an alanine, which is minimal in fact).

The free energy lost upon mutation allows to classify the residues as hot-spots ($\Delta\Delta G_{bind} \geq 4$ kcal/mol), warm spots (4 kcal/mol > $\Delta\Delta G_{bind} \geq 2$ kcal/mol), and null spots ($\Delta\Delta G_{bind} < 2$ kcal/mol). For the sake of finding the regions determinant for binding only the hot-spots really matter. Each hot spot decreases the equilibrium binding constant for 3 orders of magnitude at least. Warm spots are interesting as well if they are in the vicinity of hot spots and can be occluded by a drug altogether.

The calculated $\Delta\Delta G_{bind}$ values are very similar for the equivalent residues of each chain, with only very few exceptions. In more than 83% of the cases, this difference is less than 1.5 kcal/mol, underlining the accuracy of the method.

The residues classified by cASm protocol as hot- and warm-spots are not located in a specific area, but rather evenly distributed along the contact surface of the dimer (Figure 46). From all the 58 residues considered for each dimer of HMG-CoA-R, only 10 residues have an energetic contribution higher than 4 kcal/mol to the formation of the dimer and therefore classified as hot-spots. These are Glu528, Ile531, Met534, Tyr644, Glu665, Asn686, Lys692, Lys735, Met742 and Val863. They are the ones that are determinant to the dimerization process and therefore they are of most importance for the development of drugs that are able to prevent or disrupt dimer formation, and hence inhibiting HMG-CoA-R enzymatic activity. The residues that have a $\Delta\Delta G_{bind}$ value between 2 and 4 kcal/mol are less important but still vital to the binding process and are designated as warm-spots. In this work a total of 25 residues were classified as warm-spots (Lys499, Leu503, Tyr511, Leu512, Tyr517, Cys527, Tyr533, Ile536, Val538, Val540, Thr558, Glu559, Arg595, Leu681, Val683, Tyr687, Lys691, Glu730, Ile746, Ile762, Asp767, Asn771, Val772, Gln819 and Met820). The remaining residues are classified as null-spots and their contribution for the dimer formation/stabilization is very small (below 2 kcal/mol). The large number of warm spots that were obtained in this work

may seem odd when compared to that of null spots, but we have to remember that this analysis was done only to a subset of the residues present on the dimer interface of HMG-CoA-R and the selection these residues was made in order to decrease the number of null spots as much as possible without excluding the warm and in particularly the hot spots residues. It is worth to note that the location of the hot and warm spots are in agreement with previous studies that indicate those same regions as important for the dimerization process[17].

Even though, for the great majority of the case, the four $\Delta\Delta G_{bind}$ energies obtained for the four corresponding residues in each subunit is equivalent, there are five that present some variability. Those residues are Tyr479, Glu528, Glu665, Ile746, and Val772.

In the case of Glu528 the $\Delta\Delta G_{bind}$ values calculated for chain B, C and D vary from 12.2 to 17.8 kcal/mol. Even though it seems a huge discrepancy, they are nonetheless much higher than the 4 kcal/mol needed for a residue to be considered a hot-spot. The problem is in the result obtained for the residue located in chain A, which has a value of 3.0 kcal/mol. This is much lower than the other values, and below the limit of energy to be considered a hot-spot. However, looking in detail to the tridimensional structure of the model, we can easily see that Glu528 of chain A is a lot more exposed to the solvent than those of the other subunits. The $SASA_{dimer}$ values reflects this issue and shows that area exposed to the solvent is much smaller in monomer A than in the other monomers (73.1 Å² versus at chain 107.2 Å² in chain B, 117.2 Å² in chain C and Å² 133.0 Å² in chain D). This happens because Glu528 from monomer A interact with the C-terminus of monomer B that is incomplete in the x-ray structure and lacks more residues in that region than those of the other monomers. This renders Glu528 from monomer A to be less shielded from the solvent, and therefore it presents lower $\Delta\Delta G_{bind}$ values. Taking this into account, we opt to exclude the $\Delta\Delta G_{bind}$ value of Glu528 from monomer A and classified it as an hot-spot taking into account the average of the $\Delta\Delta G_{bind}$ values of monomer B, C and D.

A similar case happens with Tyr479 from monomer D. For monomers A, B and C we got a $\Delta\Delta G_{bind}$ of 2.7, 2.1, and 2.4 kcal/mol, respectively, which classifies this residue as a warm-spot. However, for Tyr479 from monomer D the result was 0.4 kcal/mol that is a value characteristic of null-spots. Looking again to the amino acid sequence of the enzyme, we can see that C-terminal of monomer D starts only on residue 477, while monomers A, B and C start on residues 462 (for both A and B) and 468 (for C).

Table 3 - Differences in the $\Delta\Delta G_{binding}$ for each of the 232 mutated residues. The hot-spots ($\Delta\Delta G_{bind} \geq 4$ kcal/mol) are marked red, the warm-spots ($\Delta\Delta G_{bind}$ between 2 and 4 kcal/mol) are marked yellow, and the null spots ($\Delta\Delta G_{bind} < 2$ kcal/mol) are marked white.

Mutated Residue	$\Delta\Delta G_{binding}$ Subunit A (kcal/mol)	$\Delta\Delta G_{binding}$ Subunit B (kcal/mol)	$\Delta\Delta G_{binding}$ Subunit C (kcal/mol)	$\Delta\Delta G_{binding}$ Subunit D (kcal/mol)	Average $\Delta\Delta G_{binding}$ (kcal/mol)	Range* (kcal/mol)
Tyr479	2.7±0.9	2.1±0.9	2.4±0.9	0.4±0.9**	2.4	0.3
Lys480	-0.9±0.9	-1.0±0.9	-1.2±0.9	-0.8±0.9	-1.0	0.1
Leu499	2.4±0.9	2.7±0.9	2.6±0.9	2.8±0.9	2.6	0.1
Lys502	0.8±0.9	-0.1±0.9	-0.4±0.9	0.7±0.9	0.3	0.6
Leu503	2.2±0.9	2.3±0.9	2.2±0.9	2.3±0.9	2.3	0.1
Ser508	0.8±0.9	-0.1±0.9	0.9±0.9	0.4±0.9	0.5	0.4
Tyr511	2.8±0.9	2.1±0.9	2.2±0.9	2.2±0.9	2.3	0.3
Leu512	2.7±0.9	2.6±0.9	2.4±0.9	2.6±0.9	2.6	0.1
Tyr517	2.0±0.9	1.8±0.9	2.3±0.9	2.0±0.9	2.0	0.2
Cys526	0.3±0.9	0.6±0.9	1.4±0.9	0.5±0.9	0.7	0.4
Cys527	3.1±0.9	3.0±0.9	3.2±0.9	2.7±0.9	3.0	0.2
Glu528	3.0±0.9**	12.2±0.9	17.8±0.9	14.0±0.9	14.7	5.6
Asn529	0.4±0.9	0.0±0.9	0.5±0.9	1.4±0.9	0.6	0.6
Ile531	5.3±0.9	4.7±0.9	4.5±0.9	5.2±0.9	4.9	0.3
Tyr533	3.0±0.9	2.9±0.9	3.0±0.9	2.9±0.9	3.0	0.1
Met534	5.3±0.9	6.3±0.9	5.9±0.9	5.1±0.9	5.6	0.5
Ile536	3.8±0.9	3.6±0.9	3.8±0.9	3.9±0.9	3.8	0.1
Val538	3.3±0.9	4.1±0.9	3.1±0.9	3.4±0.9	3.5	0.4
Val540	2.9±0.9	4.0±0.9	3.6±0.9	3.3±0.9	3.4	0.6
Gln552	1.6±0.9	1.5±0.9	1.8±0.9	1.8±0.9	1.7	0.1
Thr558	1.6±0.9	2.3±0.9	1.7±0.9	3.2±0.9	2.2	0.6
Glu559	2.8±0.9	3.8±0.9	3.3±0.9	5.5±0.9	3.9	1.0
Arg571	-0.7±0.9	1.09±0.9	-0.7±0.9	-1.5±0.9	-0.4	1.1
Arg595	1.9±0.9	1.5±0.9	4.7±0.9	2.3±0.9	2.6	1.1
Tyr644	4.5±0.9	3.9±0.9	3.9±0.9	4.6±0.9	4.2	0.4
Asn658	0.3±0.9	0.6±0.9	2.1±0.9	1.1±0.9	1.0	0.7
Lys662	0.4±0.9	-0.6±0.9	0.1±0.9	0.0±0.9	0,0	0.4
Glu665	2.6±0.9**	3.8±0.9**	2.9±0.9**	8.7±0.9	8.7	-
Leu681	2.0±0.9	1.9±0.9	3.0±0.9	1.4±0.9	2.1	0.6
Val683	1.8±0.9	2.1±0.9	2.0±0.9	3.6±0.9	2.4	0.5
Ser684	0.3±0.9	0.1±0.9	0.1±0.9	-0.1±0.9	0.1	0.2
Asn686	4.7±0.9	5.5±0.9	5.4±0.9	4.6±0.9	5.0	0.5
Tyr687	2.2±0.9	2.0±0.9	1.9±0.9	2.2±0.9	2.1	0.1
Lys691	2.9±0.9	2.5±0.9	4.1±0.9	2.4±0.9	3.0	0.6
Lys692	5.5±0.9	5.6±0.9	7.8±0.9	4.4±0.9	5.8	1.4
Glu730	1.6±0.9	2.1±0.9	2.4±0.9	3.3±0.9	2.3	0.8
Val731	0.5±0.9	0.4±0.9	0.1±0.9	0.6±0.9	0.4	0.2
Asn734	0.7±0.9	0.8±0.9	1.0±0.9	2.0±0.9	1.1	0.4
Lys735	3.9±0.9	4.4±0.9	4.1±0.9	3.9±0.9	4.0	0.2
Val738	1.0±0.9	1.2±0.9	1.4±0.9	1.2±0.9	1.2	0.2
Met742	4.0±0.9	4.6±0.9	4.5±0.9	3.7±0.9	4.2	0.4
Ser745	0.4±0.9	-0.1±0.9	0.2±0.9	0.2±0.9	0.2	0.2
Ile746	0.9±0.9	3.7±0.9	3.7±0.9	0.9±0.9	2.3	1.4
Asn755	0.2±0.9	1.0±0.9	0.3±0.9	0.4±0.9	0.5	0.3
Thr758	1.0±0.9	2.0±0.9	1.9±0.9	1.6±0.9	1.6	0.4
Ile762	2.5±0.9	2.7±0.9	3.1±0.9	2.4±0.9	2.7	0.3
Asp767	3.2±0.9	3.2±0.9	3.0±0.9	3.1±0.9	3.1	0.1
Asn771	1.6±0.9	2.4±0.9	2.7±0.9	1.3±0.9	2.0	0.7
Val772	4.1±0.9	2.0±0.9	2.1±0.9	4.0±0.9	3.0	1.0
Ser775	0.4±0.9	0.9±0.9	0.3±0.9	0.1±0.9	0.4	0.3
Leu812	1.3±0.9	1.4±0.9	1.6±0.9	1.1±0.9	1.3	0.2
Gln814	1.4±0.9	0.6±0.9	1.0±0.9	1.3±0.9	1.1	0.3
Gln819	1.5±0.9	2.3±0.9	2.1±0.9	2.3±0.9	2.0	0.3
Met820	3.5±0.9	4.2±0.9	4.8±0.9	3.3±0.9	3.9	0.6
Leu857	2.4±0.9	2.1±0.9	0.9±0.9	1.8±0.9	1.8	0.6
Leu862	1.5±0.9	1.2±0.9	2.6±0.9	1.1±0.9	1.6	0.5
Val863	4.0±0.9	4.5±0.9	9.3±0.9	4.4±0.9	5.5	1.6
Lys864	-0.5±0.9	-0.2±0.9	-0.5±0.9	-0.9±0.9	-0.5	0.3

*The range of a given set of values is defined as the difference between the higher and the lower values of the set.

**This residue was not taken into account in the calculation of the average $\Delta\Delta G_{bind}$ as explained in the text.

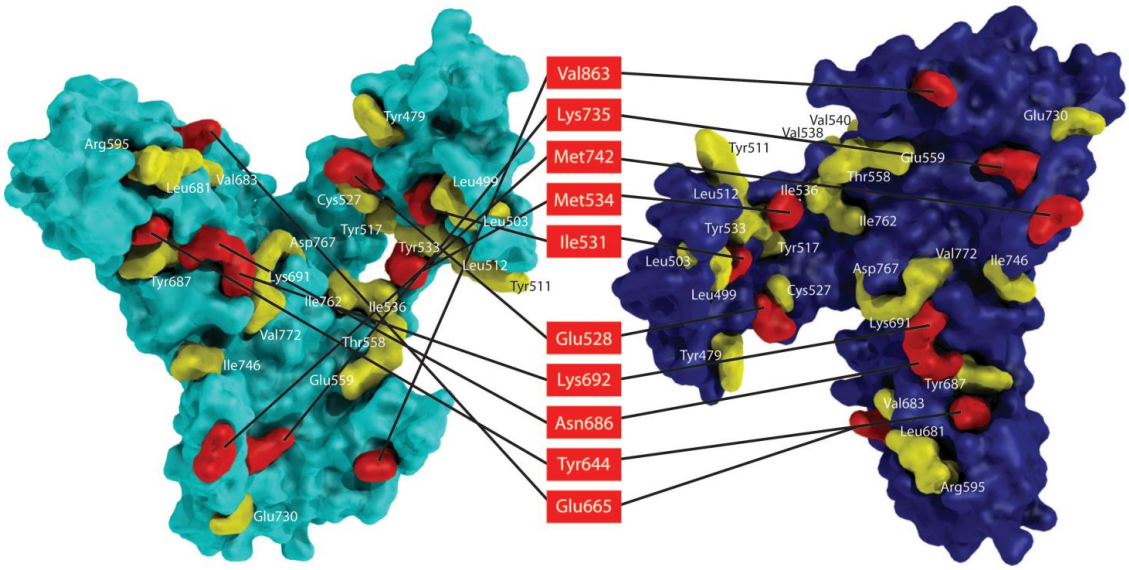


Figure 46 – "Open book" representation of the two monomers with the location of the warm and hot-spots in the monomer in the dimer CD. Hot-spots are represented in red and warm-spots in yellow. Each residue are classified as either a hot or warm spot by taking into account the average values of the $\Delta\Delta G_{bind}$.

Accordingly, Tyr479 from monomer D is not so well shielded from the solvent as its counterparts from the other monomers. The $SASA_{dimer}$ values reflect this trend 71.5 Å² for chain A, 70.5 Å² in monomer B, 60.2 Å² in monomer C and 53.0 Å² in monomer D). This means that the $\Delta\Delta G_{bind}$ of Tyr479 from monomer D can be excluded since it is a result by the lack of the full sequence of monomer in that region. We have therefore classified Tyr479 as a warm-spot taking into account the average value of the $\Delta\Delta G_{bind}$ values of monomer A, B and C.

The $\Delta\Delta G_{bind}$ values of Glu665 from monomer D are also different from the ones obtained for the other monomers. The value from monomer D classifies this residue as a hot-spot and the other three as warm-spots. Looking at the three-dimensional structure of the full protein it is evident that these differences come from the location of the Glu665 in each monomer. In all cases this amino acid residue is located very close to the N-teminus and, as it was described before, this part of the monomer finishes at different residues. Indeed, Glu665 from chain D is more deeply buried in monomer C than its counterparts of the other subunits, which lack more residues in that region. The calculated $SASA_{dimer}$ values upon dimerization reflect this trend. In monomer D, Glu665 has a $SASA_{dimer}$ value of 79.2 Å², while in chains A, B and C this values are from 62.1 Å², 51.4 Å² and 45.6 Å², respectively. Based on this information, we only took in account the result that was obtained with chain D and therefore classify this residue as a hot-spot.

Ile746 and Val772 are two other residues that also show different $\Delta\Delta G_{bind}$ values on the four monomers. These residues are located in the center of the tetramer in the only region where the four monomers interact with each other. The calculated $\Delta\Delta G_{bind}$ values reveal that these values seem to be counterbalancing each other in order to maintain the full complex in a minimum. Therefore the higher $\Delta\Delta G_{bind}$ values of Ile746 in monomer B and C are compensated by the lower $\Delta\Delta G_{bind}$ values observed for the same residue in monomer A and D. MD simulation also emphasizes this, showing that the results are dependent on the stabilization of each residue in two different minima. In order to classify this residue as a null- or warm-spot we have made an average of the four $\Delta\Delta G_{bind}$ values. The final results classified these residues as a warm spots.

Table 4 - Properties calculated for null, warm and hot spots. Percentages were based on the total SASA for the free residue. The average hydrophilic contribution corresponds to the sum of $\Delta\Delta E_{electrostatic}$ and $\Delta\Delta G_{PolarSolv}$. The average hydrophobic contribution corresponds to the sum of $\Delta\Delta_{vdW}$ and $\Delta\Delta G_{NonPolarSolv}$.

	Hot Spots	Warm Spots	Null Spots
Residues identified in each monomer	10	25	23
Average $\Delta\Delta G^{Bind}$ (kcal/mol)	6.3	2.7	0.7
Average $\Delta\Delta_{Eelectrostatic}$ (kcal/mol)	8.6	3.3	-1.7
Average $\Delta\Delta_{EvdW}$ (kcal/mol)	5.0	3.0	1.6
Average $\Delta\Delta G_{PolarSolv}$ (kcal/mol)	-7.4	-3.7	0.7
Average $\Delta\Delta G_{NonPolarSolv}$ (kcal/mol)	0.2	0.1	0.1
Average hydrophilic contribution (kcal/mol)	1.2	-0.4	-1.0
Average hydrophobic contribution (kcal/mol)	5.1	3.1	1.7
Average SASA in Monomer ($Å^2$ and %)	102.6 (35.6)	86.6 (30.4)	82.4(30.3)
Average SASA in Dimer in ($Å^2$ and%)	8.5 (3.0)	16.7(5.9)	29.6(10.8)
Average SASA lost upon Dimeration ($Å^2$ and %)	94.0(33.1)	69.9(25.0)	52.8 (19.5)

Table 4 displays the values of a set of relevant properties calculated for null, warm and hot spots. These results show that the hydrophilic interactions that stabilize these residues in the dimer interface ($\Delta\Delta E_{electrostatic}$) are similar to the ones that stabilize the monomer alone in the aqueous solvent ($\Delta\Delta G_{PolarSolv}$). The hydrophobic contributions are, however, quite different. The hydrophobic interaction of these residues on the dimer

interface ($\Delta\Delta E_{vdW}$) is much higher than of the monomer alone with the solvent ($\Delta\Delta G_{NonPolarSolv}$). These results show that the hydrophobic interactions are the major contributor for the binding energy, and therefore the main driving force for the dimerization of HMG CoA-R.

6.3.3. SASA Analysis

Besides the $\Delta\Delta G_{bind}$ results, we have also calculated the values of SASA (Solvent Accessible Surface Area) for each residue, considering the full dimer ($SASA_{dimer}$) and considering only the residues in the same subunit ($SASA_{monomer}$). Given the large difference in size that characterizes the different residues present at the interface, the values are presented as a percentage of the potential SASA for the free residue as well. This means that a high $SASA_{monomer}$ indicates that a given residue is poorly shielded by residues from its own subunit and exposed to the solvent and/or to residues from the other subunit, while a low $SASA_{monomer}$ indicates a residue that is highly protected from other interactions by residues of its own subunit. Similarly, a high $SASA_{dimer}$ illustrates that, in the dimer, a given residue is a high exposition to the solvent, while a low $SASA_{dimer}$ indicates it is extremely protected by residues from the two subunits. From the difference between $SASA_{dimer}$ and $SASA_{monomer}$ (described as SASA lost upon dimerization), it is possible to assess how buried a given amino acid is in the surface of the other subunit. The percentage of buried area (in relation to the free residue) is also interesting because it shows how much of the "potential for interaction" is effectively used in the dimer (i.e. how much optimized is the interaction).

The results expressed in Table 4 show that all the hot-spots that were identified in this study have a very low SASA in the dimer. Upon dimerization these residues have in average only 8.5 Å² of their area exposed to the solvent, which corresponds in to only 3.0% of the total of the available area of the residue. These values contrast markedly with the estimated $SASA_{monomer}$ values for these residues that in average are around 102.6 Å² and correspond to a mean of 35.6% of the total area of the residue. Considering the $SASA_{dimer}$ and $SASA_{monomer}$ values, it becomes evident that dimerization leads to a major decrease in the SASA of these residues (in average 94.0 Å²) which, in turn, accounts for a decrease of 32.6% of the available area of the residue that now interacts with the other monomer.

The warm-spots are slightly less exposed than the hot-spots when considering only the monomer (in average 86.6 Å² versus 102.6 Å² in the hot-spots). Upon dimer formation the corresponding SASA decreases to 16.7 Å² (compared to 8.5 Å² in the hot spots),

which represents an average decrease of 69.9 $Å^2$ per amino acid residue (25.0% of the total area of the residue).

The null spots interact preferentially with their own subunit, exhibiting small $SASA_{monomer}$ values, which decrease slightly when considering the dimer (the $SASA_{monomer}$ values for the null-spots are similar values to the ones of the warm spots). Their values for the $SASA_{monomer}$ and the $SASA_{dimer}$ are 82.4 $Å^2$ and 29.6 $Å^2$, with an average decrease of 52.8 $Å^2$ on the SASA upon dimerization (versus 69.9 $Å^2$ in the warm-spots). Note than many null-spots were eliminated in the initial stage of this study, when the threshold of 40 $Å^2$ for the $SASA_{monomer}$ was applied. This is also explains why the number of null-spots is very small when compared with the number of warm- and hot-spots.

The results of the SASA analysis of this work are in line with previous proposals, which dictate that there is a correlation between the buried surface area and the free energy of binding[388, 391, 393, 396-397].

6.3.4. Druggable Sites for Dimerization Inhibitors

Two regions (region A and B in Figure 47) that have been identified as optimal to drug and preclude the formation of an active dimer of HMG-CoA-R are located in distinct regions of the monomers.

Region A is located in one extreme of each dimer and opposite to the dimer:dimer interface. This means that the hot- and warm-spots that were identified in this region are not important for the formation of the tetramer, but instead for the dimer formation. This region is also the one where a higher concentration of hot and warm spots is found, and therefore the one that might be more important for the dimerization process.

Figure 48 shows in greater detail the surface area of that region. It is very wide and shows a distinct cleft in the center where the hot spots from the neighbor monomer fit and bind. Such cleft should be an optimal druggable site for dimerization inhibitors capable of blocking the binding of the other monomer and therefore impair the formation of an active HMG-CoA-R.

This druggable pocket has a truncated cone shape with an internal volume of around 194 $Å^3$ and dimensions of around 7.6 Å X 6.3 Å X 4.9 Å (Figure 48). Such dimensions are very convenient as they can fit a molecule with 3-4 fused rings inside, or even greater if a part of the drug protrudes outside the pocket.

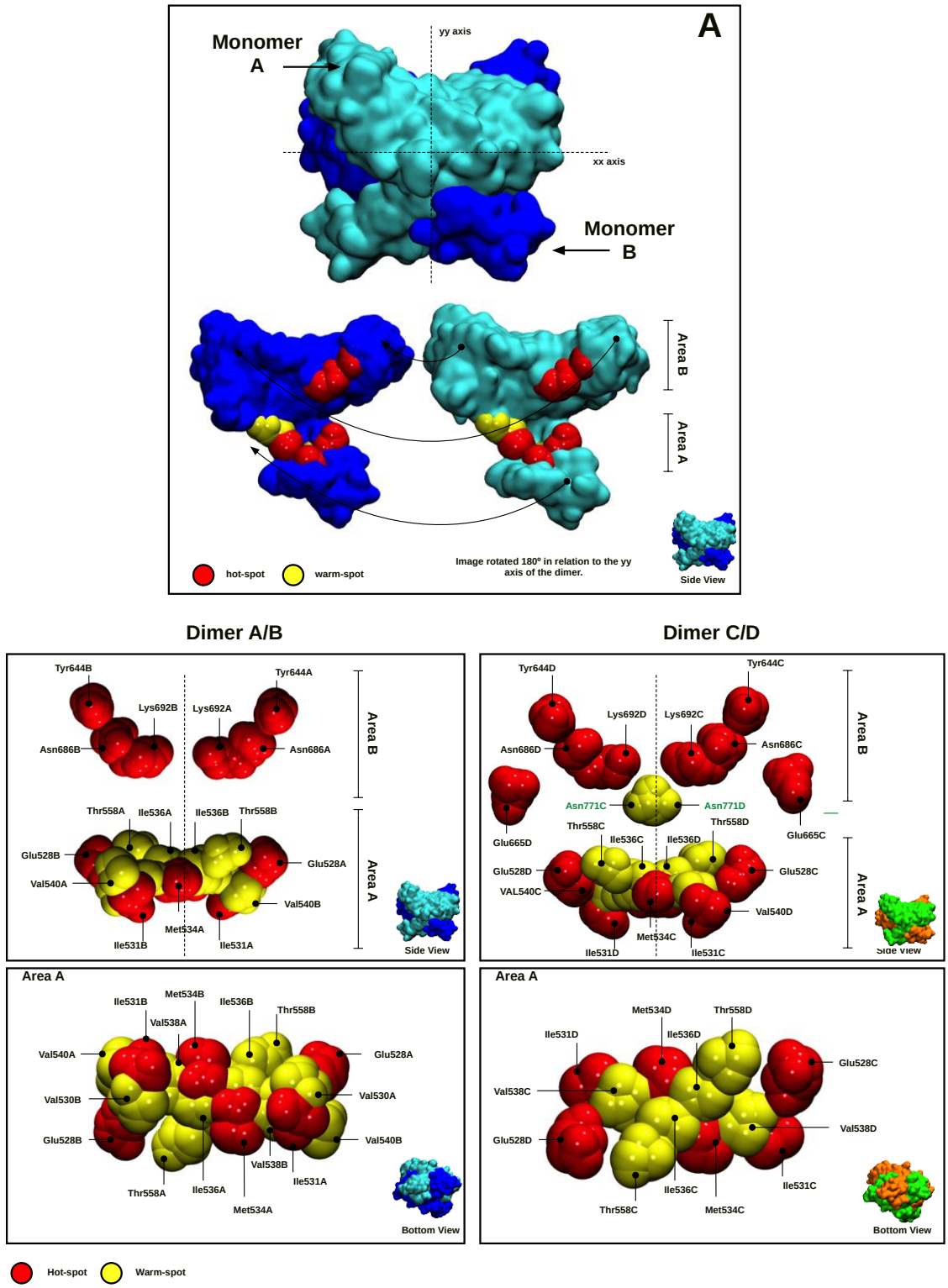


Figure 47 - Schematic representation of the interface of each dimer constitutive of HMG-CoA-R, illustrating the relative position of all the hot- and warm-spots identified by cASM protocol.

The other region that can also be used to preclude the formation of an active dimer (Region B) is located very near to the active site of the enzyme. The active site of HMG-CoA-R is composed by residues of both monomers and has a V-shape due to the

presence of two tunnels that are occupied by the two cofactors that are required for the catalytic process, i.e., CoA and NADPH. The binding pocket of HMG is located in the part where both tunnels join and it is formed by residues Ser684, Asp690, Lys691, Lys692 and Asp767 from one subunit and Glu559, Lys735, Asn755, Leu853 and His866 from the other.

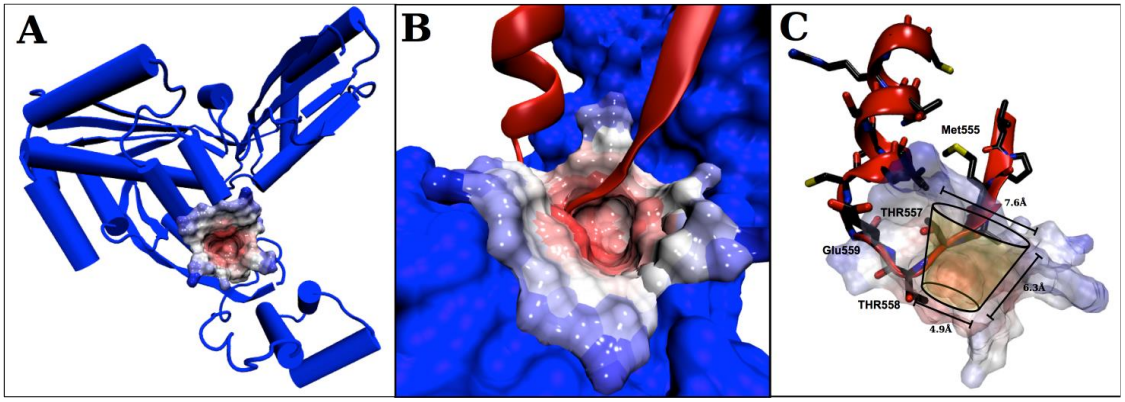


Figure 48 - A) Druggable cavity in the interface between monomers A and B of HMG-CoA-R, where the most important hot and warm spots are located. B) Top view of the druggable site. In red are represented the portions of chain B that interact with the cavity. C) Side view of the druggable cavity highlighting its shape, dimensions, and polar residues present in that region. The pockets can fit a drug of 194 Å^3 totally inside the pocket.

Taking as reference one of the active sites formed between monomers A and B (Figure 49), the hot-spots that were identified in region B are only located in monomer B and close to the binding position of CoA and HMG. One of the identified hot-spots, Lys692, is also one of the residues that is required for the catalytic process. Note that this region does not present a pocket, is mostly flat. Therefore, even being so important for dimerization, this region should be much more difficult to drug than region A above, where a perfect pocket accommodates drugs of moderate size.

The statins that are commercially available on the market to inhibit HMG-CoA, as it is for example the case of atorvastatin, bind exactly in the same region of acetyl-CoA and HMG and interact with one hot-spot that belong to region B. In addition taking as reference Figure 49, these compounds interact preferentially with just one of the monomers, in this case monomer B. These results suggests that besides the competitive inhibition of these compounds, modified versions of these inhibitors could also be used to impair the dimerization of HMG-CoA-R (by occluding the hot-spots of region B) or, at least, deviate the residues of the active site from its the correct alignment, in order to form an active HMG-CoA.

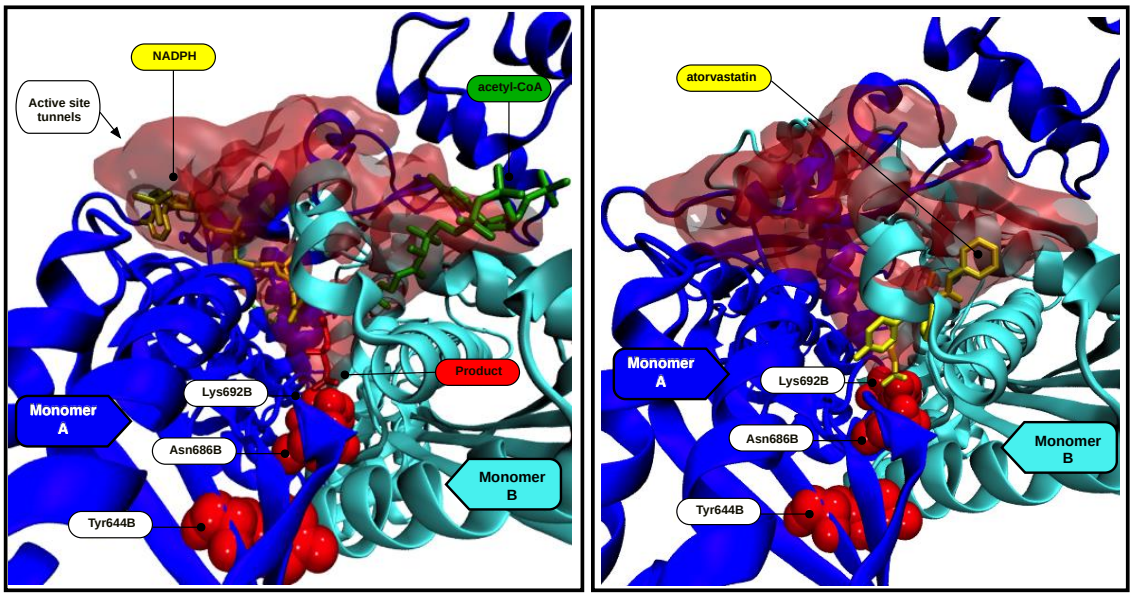


Figure 49 - Left: One of the active sites of HMG-CoA-R present at dimer interface of monomers A and B (pdb code: 1dqa). The NADPH, acetyl-CoA and the product of the catalytic process are represented in sticks in yellow, green and red respectively. The hot spots of region B are represented in van der Waals style and colored in red. Right: Binding position of one of the statins commercially available in the market, atorvastatin (pdb code: 1hwk).

6.4. Conclusions

In this study, we presented a detailed atomistic analysis of each monomer interface in the HMG-CoA-R enzyme, an important target to control the levels of cholesterol in the blood. In particular, we have evaluated the energetic contribution of all the residues at the monomer interface of this tetrameric enzyme, in an attempt to identify the most important determinants for dimer formation. This analysis was further complemented with a solvent accessible surface areas study for all interfacial residues and the decomposition of the binding energy into hydrophilic and hydrophobic contributions, where we have shown that the driving force for oligomerization is essentially hydrophobic. The outcome provides relevant clues for the development of new inhibitors of HMG-CoA-R designed to block the association of the monomers, thereby preventing the formation of the fully active dimer enzyme.

Ten residues (Glu528, Ile531, Met534, Tyr644, Glu665, Asn686, Lys692, Lys735, Met742 and Val863) were shown to be fundamental for dimer formation. Mutating one of these residues by alanine leads to a destabilization of the resulting dimer formed of more than 4 kcal/mol when comparing with the wild-type HMHG-CoA. These residues were also the largest contributors, in terms of area, to the subunit interface of the dimer (94 Å2 each, in average). Three other pairs of residues also make a significant energetic contribution (albeit more modest than the latter) to dimer formation. These residues,

identified as warm spots, are Val538, Thr558 and Ile536, and have also an intermediate contribution in terms of area to the subunit interface.

Based on the location of the hot and warm spots, these residues have been grouped in two distinct areas of the protein, that are proposed as two promising druggable targets to develop new dimerization inhibitors of HMG-Co-A-R. One of these regions is very close to the active site and closely resembles the one that is occupied by the statins that are currently commercially available to inhibit HMG-CoA. The other region is where a higher concentration of hot and warm spots has been identified in this study, has a deep pocket that fits drugs of moderate size (a volume of 194 A3), and therefore should be the one that is most easily druggable.

The inhibitors targeting these two regions should be capable of interfering or disrupting the more relevant and persistent dimerization contacts, blocking the formation of the fully active HMG-CoA-R

These features should be taken into account in future drug design and development studies targeting HMGCoA-R.

CHAPTER 7

STUDYING THE MAMMALIAN ENZYMATIC COMPLEX FAS

7.1. Fatty Acid Synthase

As stated previously, FAS is an enzymatic complex responsible for *de novo* synthesis of fatty acids. The discovery of FAS and how fatty acids are synthesized goes back more than one century ago, when Henry Stanley Raper determined in 1907 that fatty acids were likely derived from a C_2 precursor [398]. Still, at that time Raper was not able to ascertain whether this precursor was ethanol, acetaldehyde or acetic acid. It was not until almost half a decade later, in 1953, that it was found that this precursor was an activated form of acetate, acetyl-CoA [399]. A few years later, however, in 1958, two different groups [400-401] independently arrived to the same conclusion; prior to being used in the synthesis of fatty acids, acetyl-CoA was carboxylated to malonyl-CoA, which was then decarboxylated during the condensation step of the reaction.

By the late 1960s scientists were able to purify FAS from yeast and animal tissues and determined that these were different from those present in bacteria and plants[402]. At this point it was established that there should be made a difference between these two types of FAS, and so the following terminology was used:

- Type I FAS (FAS I) refers to the large multidomain enzymes that catalyze fatty acid synthesis in its entirety, disposing of the complete product only when it is complete. They are present in non-plant eukaryotes
- Type II FAS (FAS II) refers to the enzymes present in plant, prokaryotes and algae. They are expressed as individual domains and not a single protein. These domains are found on the cytosol of cells and each of them catalyzes one part of the reaction. They are encoded by several genes.

FAS I contains FAS enzymes from both animals and fungi. However it should be noted that despite belonging to the same type, these enzymes are very different is terms of structural organizations[403]. They both include the same functional domains responsible for the synthesis of fatty acids, but their disposition is quite distinctive. Fungi FAS I is a heterododecamer of 2.6 MDa, which is considerably larger than animal FAS, and is encoded by two different genes. Since there will be no further discussion about

FAS I from fungi in the remaining of this work, all subsequent mentions of FAS I will refer solely to that of animals.

Another interesting fact is, though there are many differences between FAS I and FAS II, the reactions catalyzed by the analogous domains are quite identical, since the mechanism for fatty acid synthesis has been greatly conserved throughout evolution. Also, if one compares the structural arrangement of each FAS II domain with its counterpart on the FAS II enzyme, one could easily observe that these share a great deal of resemblance (Figure 50). It is thought that FAS I is derived from an ancient gene fusion event of type I proteins. This would in part explain the similarities that exist between both types.

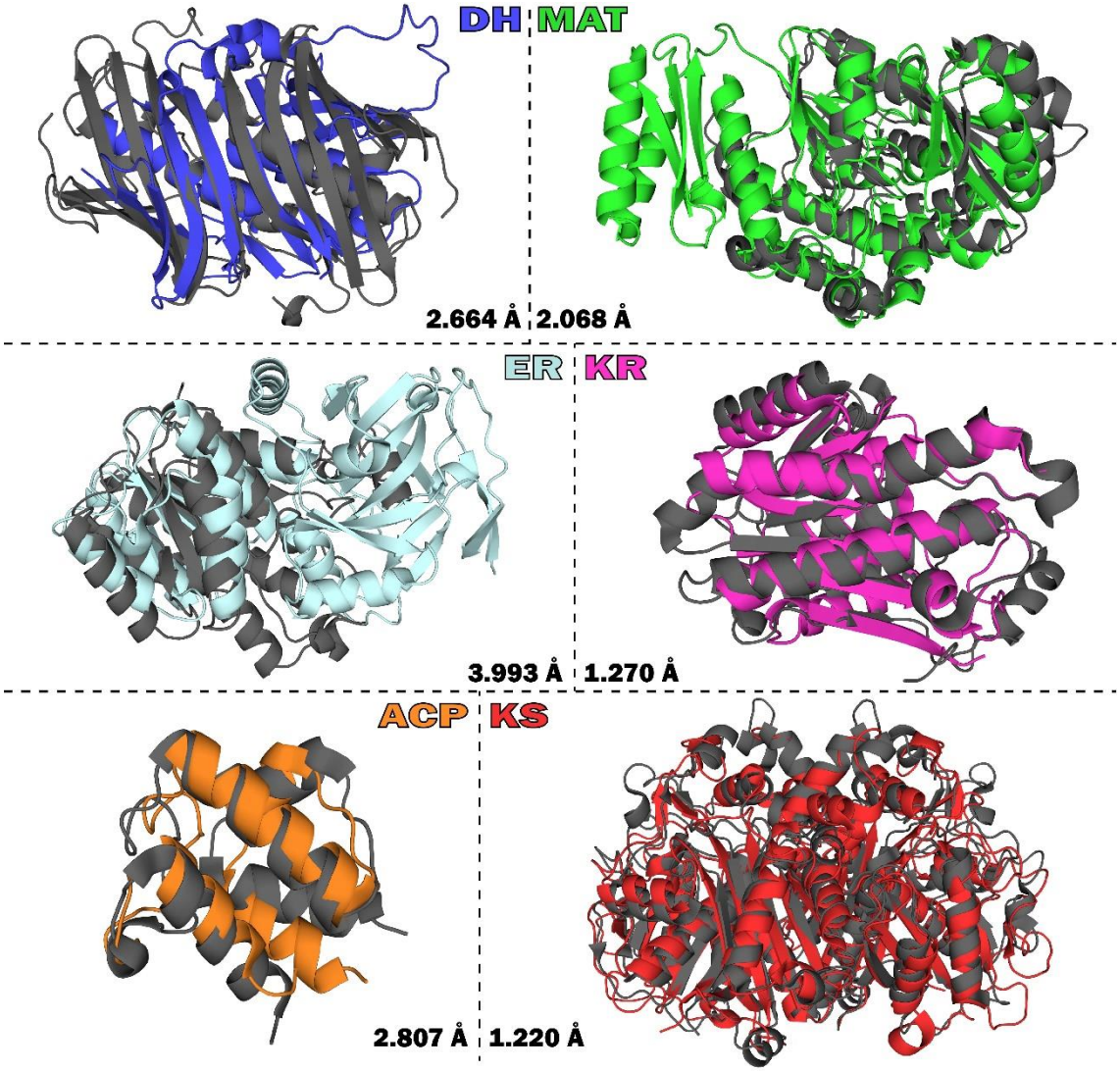


Figure 50 – FAS domains from both type one and two aligned in a way that is easy to see that the structures are quite conserved. Each type I FAS domain is colored, whereas the type II FAS domains are in grey. The numbers show the RMSd between both structures.

Mammalian FAS I (mFAS) is a multidomain homodimeric protein of approximately 540 kDa. mFAS comprises all the necessary domains to produce a single molecule of fatty acid from its building blocks (malonyl and acetyl). In total, mFAS is composed by eight different domains:

- Six catalytic domains:
    - β-Ketoacyl Synthase (KS)
    - Malonyl/acetyltransferase (MAT)
    - β-Hydroxyacyl Dehydratase (DH)
    - Enoyl Reductase (ER)
    - β-Ketoacyl Reductase (KR)
    - Thioesterase (TE)
- One carrier domain:
    - Acyl Carrier Protein (ACP)
- One structural domain:
    - Pseudo-methyltransferase (ΨME)

ACP is perhaps one of the most interesting modules in this whole complex. It is a very small, heat stable protein with a structure that is rather conserved throughout different organisms[404]. It consists of a four helix bundle which is stabilized by hydrophobic interactions that form between different helixes. In Figure 50 the RMSd between mFAS and FAS II from *E. coli* appears to be rather large (2.807 Å) but if one looks at the structure it is possible to notice that this difference is due mainly to the fact that ACP is very flexible and has several loops capable of adopting varied conformations. In fact, it is easy to perceive that the overall secondary structure of ACP is conserved between the two structures.

ACP is responsible for the transport of all the substrates, intermediates and the final product since the beginning and until the end of the reaction[405-406]. It starts by carrying the initial acetyl group on the first step of the reaction to the active site of KS and finishes by transporting the completed fatty acid to TE, were it is released. ACP's function is dependent on a prosthetic group called phosphopantetheine (PPT), which is covalently bound to a conserved serine residue present on helix 2. PPT is a cofactor derived from coenzyme A and is added to ACP post-translationally by a transferase

enzyme, and only then is it ready to perform its role in FAS. Given that ACP is such a mobile and flexible unit, it is difficult to study its structure using crystallography. As of now, the only structure available on PDB of ACP was obtain through the co-crystallization of it with the phosphopantetheine tranferase enzyme (PDB code: 2CG5[407]).

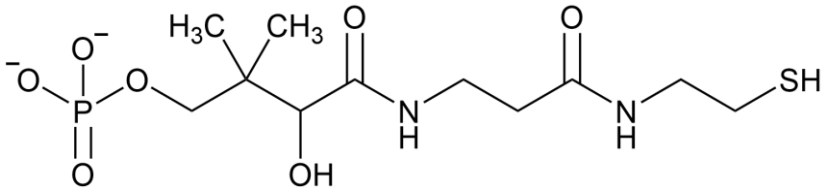


Figure 51 – Structure of the phosphopantetheine molecule

In the ACP of FAS II, helix 2, where the PPT is bound, is composed of a cluster of highly conserved and negatively charged amino acids which are regarded as an universal recognition helix responsible for the interaction with other enzymes[408]. In FAS I however the residues in this helix are not as conversed and do not retain the same negative character. Nonetheless, there is still evidence that this helix plays an important role in the docking of ACP with the other domains[409]. The internal pocket of the ACP lined with apolar residues responsible for interacting and stabilizing the acyl intermediate during transport. Since the intermediate can be as big as 16 carbons, it would be difficult to carry it through the aqueous medium, so it enters the ACP module and interacts with the hydrophobic interior, while the PPT moiety stays on the outside and in contact with the solvent.

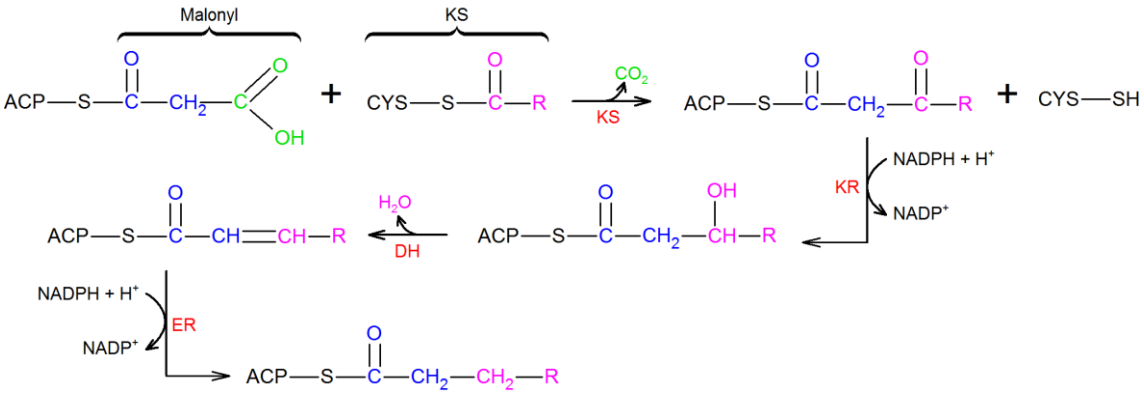


Figure 52 – Overall reaction catalyzed by FAS, with the indication of all modules in which each part occurs.

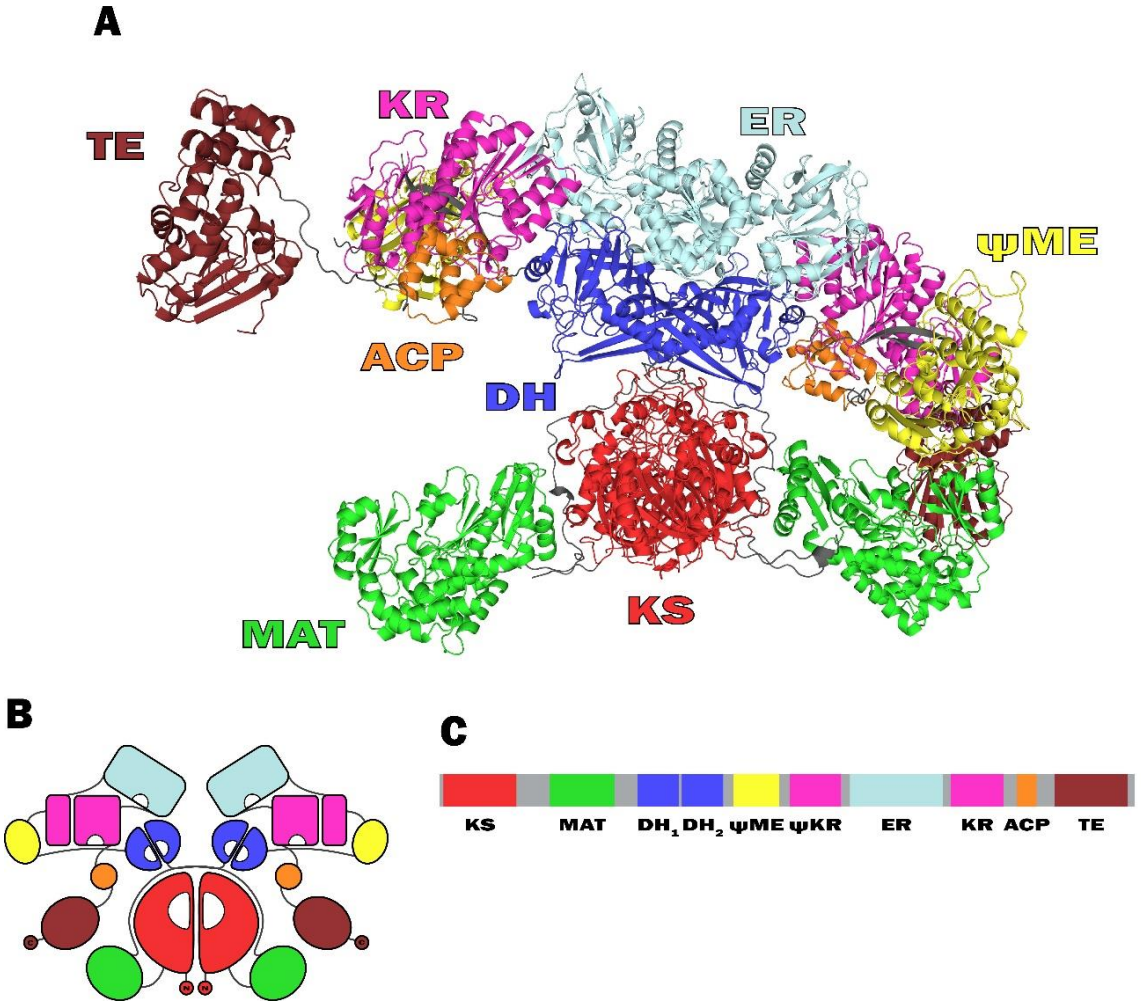


Figure 53 – Type I FAS. A: the overall structure of the type I FAS megacomplex, with all modules represented in diferent colors (KS – red, MAT – green, DH – dark blue, ER – light blue, KR – pink, ψME – yellow, ACP – orange and TE – brown). B: a schematic representation of all the subunits and how they are connected. C: representation of the primary structure of the FAS enzyme.

In order to perform its role, ACP must transport each acyl intermediate to the right domain so that the reaction can proceed[410] (Figure 52). After PPT is bound to the serine residue, ACP goes to the MAT domain, where either an acyl or malonyl moiety is attached, depending on whether it is the beginning of a new fatty acid synthesis cycle or it is already partially formed. In the case it is the start of a new cycle, acyl is bound to ACP, which then proceeds to the KS domain, where it transfers the acyl to a cysteine residue. Subsequently, ACP leave KS and returns to MAT so that a malonyl molecule can be added to PPT. This malonyl moiety is the fatty acid building block, and will add two new carbons to the growing chain. ACP goes then back to the KS domain, where the intermediate that is already attached to this domain will condensate with the malonyl adding two more carbons to the chain and releasing CO_2. Next, ACP leaves KS again and is transferred to the active site of the KR domain, where its β-keto group is reduced

by a NADPH co-factor. After leaving KR, ACP travels to DH, where the growing fatty acid is dehydrated and later follows to the ER domain for the last step in this cycle, which is a new reduction. At the end of this cycle the PPT has attached to it a fully saturated acyl intermediate, which is then transferred again to the KS domain, so that ACP is free to go to MAT and get a new malonyl moiety to condense and a new cycle can then begin. Typically, this steps are repeated for a total of 7 turns in order to build a 16 carbon fatty acid. When it is finished, ACP travels with the final product to the chain-terminating TE domain, where the fatty acid is released from PPT to the cytoplasm and a new formation cycle can then commence[410].

In FAS I (Figure 53) it is currently not known exactly how the ACP domain bearing a particular acyl intermediate is able to locate the appropriate domain for the reaction to continue. During transport, the fatty acid intermediate is carried in the hydrophobic pocket of ACP, while PPT, given its polar nature, can easily travel outside of the carrier, and stay in contact with the cytosol. It is thought that the way ACP knows it has arrived at the right domain for the next step is through the exposure of the acyl thioester bound between PPT and the intermediate[411]. Given that the reactions occur near this bond, it could be hypothesized that its presentation to the other domain (maybe accompanied with some degree of change in the conformation of ACP) is what signals ACP that it has arrived to the correct destination. If the bond presented and/or the conformation of ACP is not the correct one for that domain to carry on the reaction, then ACP detaches from it and goes in search of the right one. In theory, this approach seems to make it easier for ACP to know whether or not it is the correct domain, without having to insert the intermediate in the active site, and instead providing cues that facilitate recognition by the appropriate enzyme. However, one cannot exclude the hypothesis that the finding of the right domain is done through a random search, in which the reaction proceeds only when a certain intermediate finds the correct enzyme and the next step can occur.

7.1.1. KS Domain

The KS domain is where the first step of the synthesis of fatty acids occurs. This is where the growing fatty acid chain is elongated, through the condensation of successive molecules of malonyl.

The structure of KS is reminiscent of enzyme belonging to the thiolase superfamily. The KS domain is homodimeric, meaning that it is constituted by two identical subunits. Each subunit is constituted by two subdomains, each of which exhibits a similar

βαβαβαββ topology, which suggests that these subdomains may have evolved through an ancestral gene duplication event.

One interesting fact is that KS is essential for the correct dimerization of FAS[412]. KS has an extensive dimeric interface which is largely stabilized by hydrogen bonds formed between the two subunits. Even though domains like ER and DH also contribute in the dimerization process, forming dimers between their subunits, it has been found that they are sufficient to promote the correct folding of FAS. In fact, FAS engineered without KS is unable to form dimers, and is found entirely as a monomer[412]. On the other hand, however, in enzyme analogous to FAS (polyketide synthases) lacking both ER and DH domains, it was observed that they can dimerize, and the interface between both monomers is constituted mostly by in the interface between the two KS domains and an additional N-terminal coiled-coil appendage.

As stated previously, KS is responsible for the elongation of the fatty acid chain. The reaction mechanism proposed for KS is shown in Figure 54[413]. The residues of the enzyme important for the reaction are one conserved cysteine and two histidines that help with the stabilization of the intermediates. Before the reaction can even start, an ACP module loaded with one of the acyl intermediates (or in the case of this is the very first step of the synthesis, an acyl moiety) must enter the active site of KS through Given that the intermediate can be a very long hydrophobic chain, the interior of the KS, past the active site, forms a tunnel lined with apolar residues, which help accommodate the intermediate during the reaction. Once the substrate enters the KS domain, the reaction can then start. The first step is the transference of the intermediate from the PNS to the cysteine residue. This step is quite straightforward: the negatively charged cysteine residue attacks the carbon from the fatty acid intermediate that is connected to the Sulphur from PNS, which causes the C-S bond to break and a new bond between cysteine and the intermediate to form (Figure 54 – 1a). This in turn renders the ACP free to leave the active site and go to the MAT domain, where a new malonyl chain extender will be attached to its free end (Figure 54 - 1b).

Afterwards, ACP travels back to the KS domain, where the PNS with the malonyl again enters the active site and the reaction is then able to proceed. In addition to the malonyl, a water molecule is also needed for the second step. The water is activated by a nearby residue, most likely one of the active site's histidines, which functions as a base and abstracts a proton from the water, and its oxygen reacts with the terminal carbon from malonyl, forming bicarbonate (HCO_3^-), which dissociates from the rest of the substrate (Figure 54 - 2). A carbanion is formed in the meantime, and it is easily stabilized by the two active site histidines.

The final steps comprise the condensation per se. First, the carbanion attacks the carbon adjacent to the sulfur from the cysteine, forming a covalent bond between them (Figure 54 - 3). Subsequently this carbon dissociates from the cysteine (Figure 54 - 4) and what began as a saturated fatty acid intermediate has now grown in two more carbons and must endure other significant chemical reactions (two reductions and one dehydration) in order to become again a fully saturated moiety.

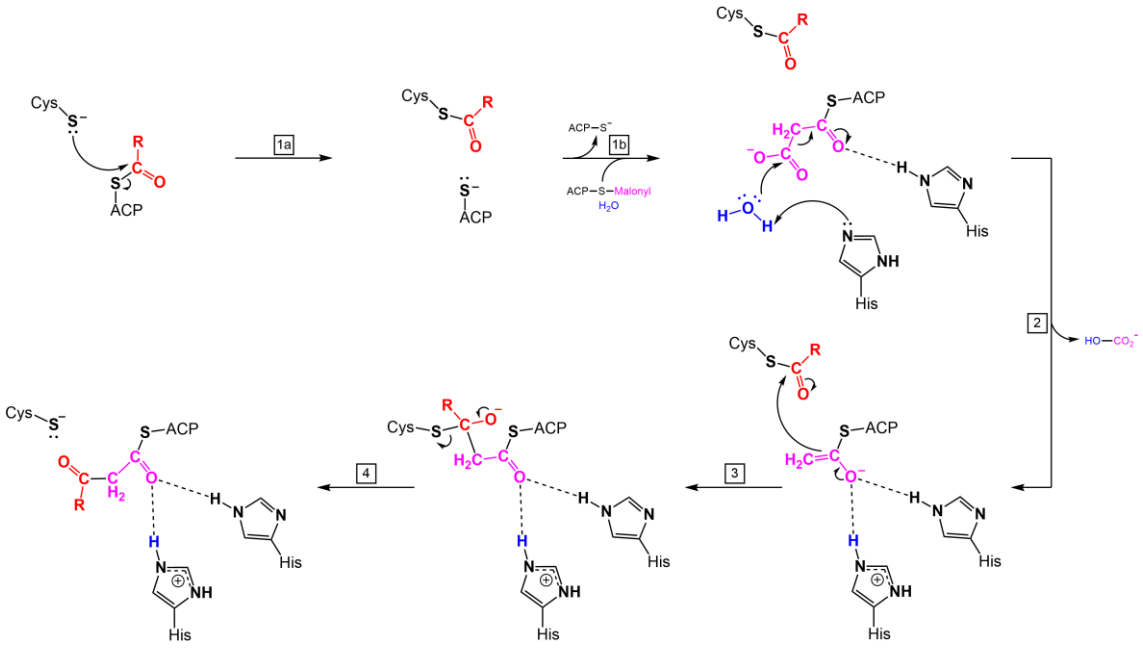


Figure 54 – Proposed reaction mechanism for the KS subunit of FAS.

At the end of all this steps, ACP detaches from KS and proceeds for the next FAS domain. To KS all that remains is to stay put until ACP arrives again with a new intermediate that needs to be condensed with one more malonyl molecule.

As one can easily see, FAS is indeed one mysterious and peculiar enzyme. Not only are intermediates shifted around in a private carriage (ACP), but also all reactions are catalyzed by the same enzyme, in different domains. Such an enigmatic protein deserves to be further studied and understood. Still, in order to study all aspects of FAS, one must choose to start somewhere and so it was decided that it would start with the KS domain, more precisely, with the study of its reaction mechanism using computational methods.

7.2. Methods

7.2.1. Making the Model

The first step in almost any computational study of an enzymatic reaction is building a model of the object of study that is able to mimic the enzyme in the most accurate way possible. In order to achieve this, the tri-dimensional structure of the protein must be obtained.

The PDB website was searched in order to find a good structure from which the KS model could be built. Since in a parallel work we had already modeled the substrate of KS in its active site, starting from the *Sus scrofa* (pig) FAS[414], and based on the similarity that exists between the pig and human forms of the enzyme, it was decided to use the KS domain from pig FAS. Meanwhile we found that the human form of FAS was also available on PDB. However as we had already a validated enzyme:substrate complex of KS, we decided to continue the study with this model, keeping in mind the similarity between both forms.

In order to build an accurate model for the study of the mechanism, the substrate had to be inserted in the active site. Given that the complete FAS structure is too big, and that the domains work independently from one another, it was decided that instead of using the complete structure, it was better to make a model of only the KS domain.

Since the substrate was not present in the FAS PDB structure, it had to be added before to the beginning of the study of the mechanism. In order to achieve that we employed a docking procedure, using the VINA software[415]. The docking was made using a rigid protein and flexible substrate. As substrate we decided on using a 10 carbon fatty acid chain attached by the carboxylic group to the PPT. We decided on a larger chain than the initial substrate (acetyl) so that we could also check how it fit in the cavity after the active site. The box where the poses were generate comprised the residues from the active site as well as those that formed a tunnel which led the PPT moiety from the surface (where it should be bound to the ACP module) to the active site.

After the docking procedure, we chose the pose with the best score to perform a MD simulations. We did a 10 ns MD simulation to see whether the substrate would stay in place or move away from the initial position. The protocol used for this simulation was similar to the one described in Chapter 6, section 6.2.1. The substrate used in the MD simulations was geometry-optimized and its electrostatic surface potential was derived at the HF/6-31G* level of theory. Then, it was parameterized with the GAFF force field[52], and its atomic charges were derived using the RESP method[56].

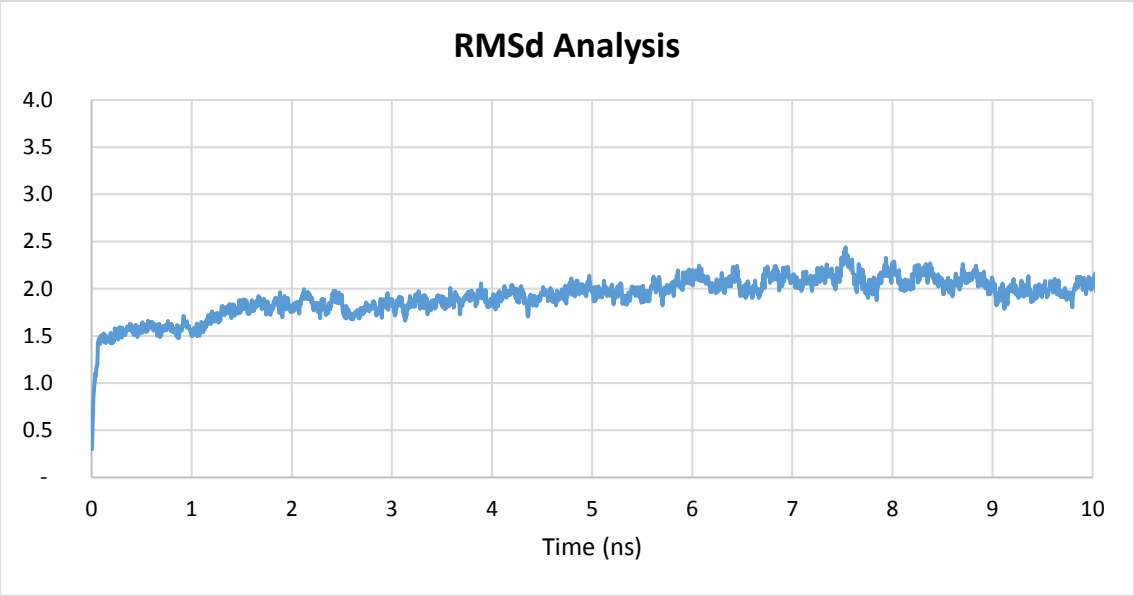


Figure 55 – Graphical representation of the RMSd values for the backbone of the model. The measurement is presented in Å.

7.2.2. ONIOM Calculations

The MD simulation showed that the substrate stayed practically in the same spot for the duration of the run. We calculated the RMSd for the whole enzyme and the substrate and concluded that after about 6 ns the enzyme was stabilized (Figure 55) and so was the substrate (in average its RMSd was below 2 Å). For the ONIOM model, we extracted a structure from the dynamics after the enzyme was stable and in which the substrate seemed in a favorable position for the reaction to occur. The substrate attached to PPT was cut so that only a molecule of acyl was in the active site. For the high layer, we selected residues His331 and His293, Cys161 and Ala160 from the active site and from the PPT-acyl we used all the acyl group, the sulfur atom and some other atoms following the sulfur (Figure 56). We did not include all of the PPT group because we found that its inclusion created some problems with the calculation and we thought it might not make a lot of difference in the final result. We used the link atom approach[65, 416-417] to saturate the valences that resulted from the truncation of bonds across the DFT and MM layers. The DFT layer was constituted by a total of 62 atoms and presented an overall charge of -1 and singlet spin multiplicity, while the overall model (total of 12233 atoms) presented a total charge of -17.

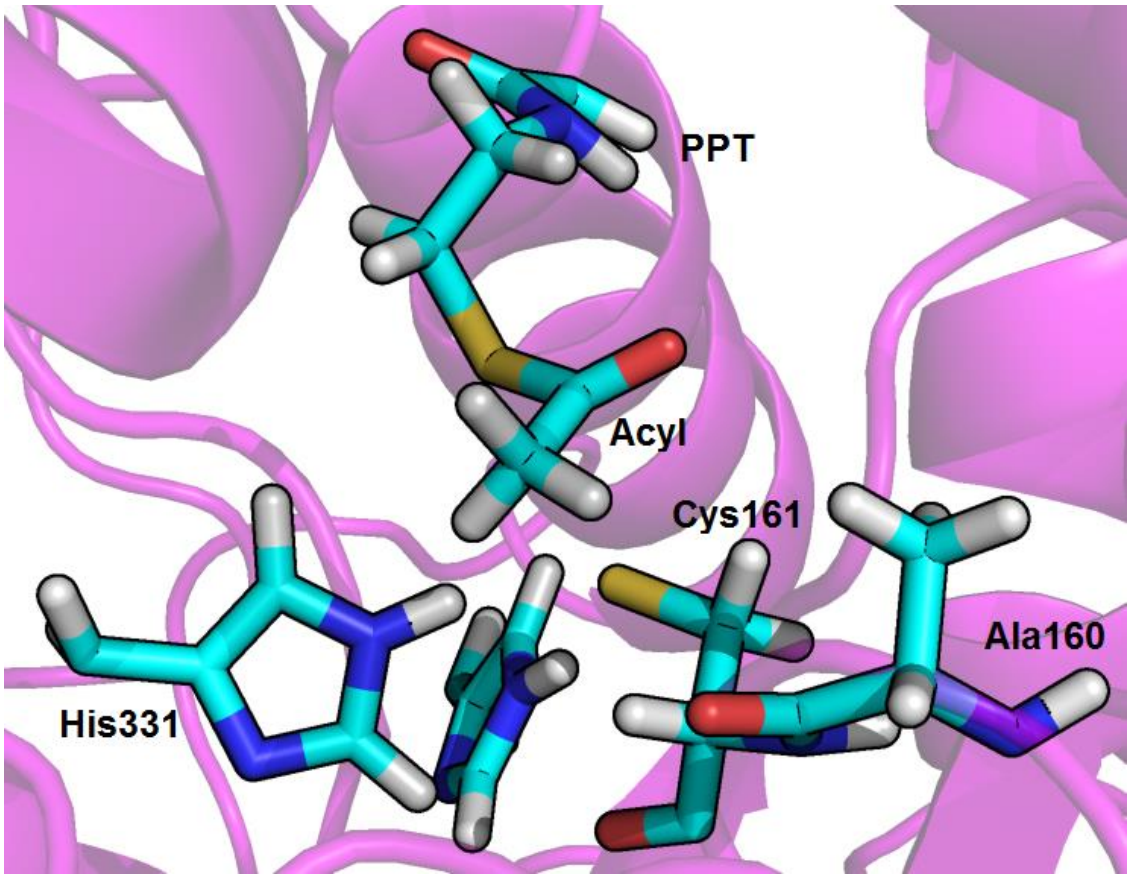


Figure 56 – High layer of the ONIOM model.

All calculations were performed with the ONIOM methodology and the electrostatic embedding scheme[67], as implemented in the Gaussian09 package[366]. Throughout the calculations all atoms farther than 12 Å from the high layer were kept frozen. The B3LYP/6-31G(d):FF99SB level of theory was used for all geometry optimizations. Several works have shown that the 6-31G(d) basis is adequate for the purpose[418].

We performed relaxed geometry optimization calculations for all stationary points along the reaction coordinate. Nuclear vibrational frequencies revealed their nature (one imaginary frequency in the DFT layer for transition states, and no imaginary frequencies in the DFT layer for minima).

7.3. Results: Step 1

As of today, this work is still in its early stages, however, we have already finished the first step of the catalytic mechanism of the KS enzyme.

In the optimized geometry we observe that the –NH of both histidine is pointing towards the sulfur atom of the Cys161, which is deprotonated. In the proposed mechanism for KS, the sulfur (S1) from the cysteine is predicted to bind to the acyl group through the carbon (C1) through which it is connected to PPT. in order to test this hypothesis we shortened the distance from S1 of the cysteine to the C1 of the acyl group and scanned this reaction coordinate.

We obtained a maximum of energy at the distance of 2.0 Å, and tested this geometry to assess whether or not it was a transition state by calculating the nuclear vibrational frequencies. This structure showed only one imaginary frequency (*i*74.5 cm^{-1}) which corresponded to a large motion of the C1 carbon of the acyl group towards the sulfur of Cys161.

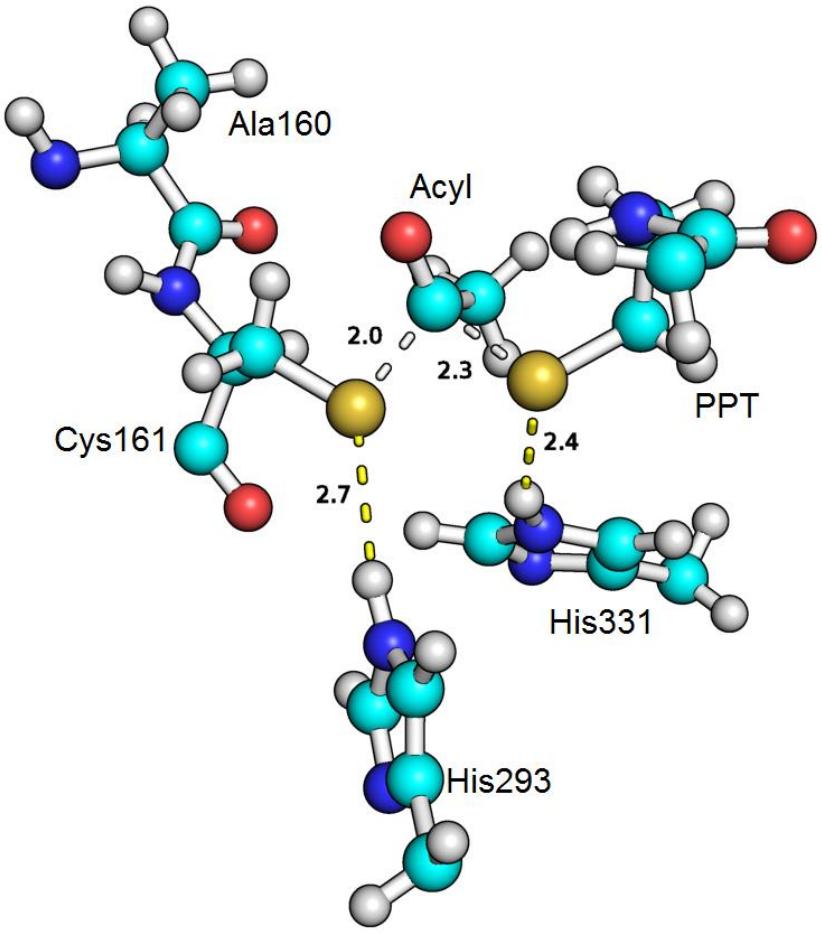


Figure 57 – Structure of the trastition state for the first step if the reaction, with distances between the significant atoms represented (in Å)

Along the reaction coordinate, we could see that as C1 and S1 got closer, the distance between C1 and the sulfur of the PPT (S2) moiety began to lengthen. It was

also clear that His293 shifted its position to stabilize the increasing negative charge of S2 (distance changed from 4.7 Å to 2.7 Å). In the transition state, it is interesting to notice that the carbon from the acyl group is midway between both sulfurs (Figure 57) and that we have one histidine stabilizing each of the sulfurs as well. After this middle point, the bond between C1 and S2 is broken (distance is 4.1 Å in the minimized product), and a new bond between C1 and S1 is formed (1.8 Å). Both cysteines are now forming hydrogen bonds with the negative sulfur from the PPT moiety (distance from His293 to S2 is 2.4 Å and from His331 is 2.3 Å).

The energy of activation for this step totals 3.26 kcal/mol, which means that the barrier is not very large. This is probably due to the fact that both histidines help with the stabilization of the negative charged sulfurs and allow for a good charge distribution which lower the barrier. The fact that the S1 from the cysteine was deprotonated also aids with the attack to the C1 carbon. The reaction is exothermic (-6.46 kcal/mol) which means it is favorable (Figure 58).

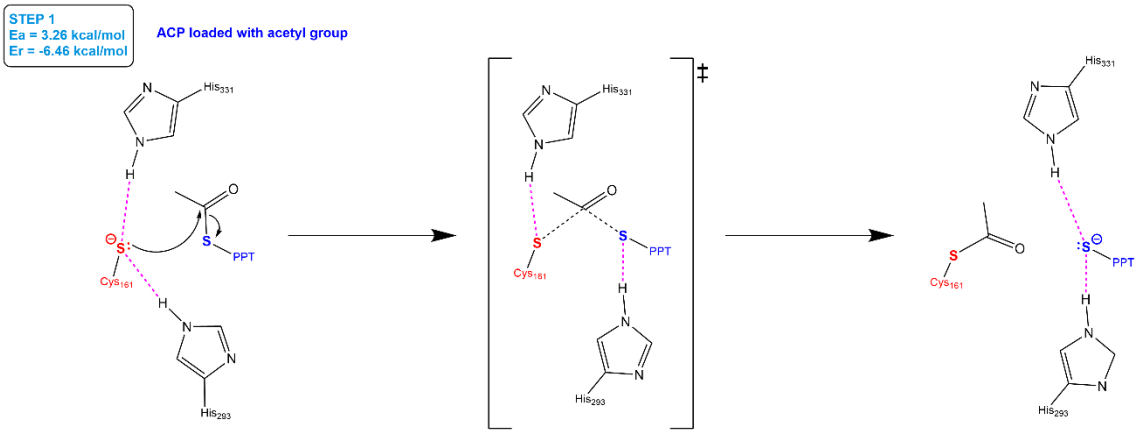


Figure 58 – The first step of the mechanism of the KS module (with representation of the transition state)

7.4. Conclusions

As of now, we still have a long way to go in order finish this work. We have currently completed the first step of the KS module of the FAS enzyme. It comprises the attack of the sulfur from Cys161 on the carbon of the acyl group which is connected to the sulfur of PPT. This step is rather straightforward and is processed according to the proposed mechanism of reaction for KS. While the acyl group travel from the PPT to the Cys161, histidines 293 and 331 also stabilize the negative charge of the sulfur atoms. In the reactant this negative charge is centered of S1 and both histidines have their hydrogens

facing this sulfur, forming two hydrogen bridges. In the transition state, however, the charge is distributed equally among both sulfurs, and so we have one histidine binding each sulfur. As for the intermediate, both histidines are now binding the same sulfur (from PPT) through a hydrogen bridge, since now the negative charge is more centered on this atom. The reaction is exothermic and it activation energy is -3.26 kcal/mol.

For future work we would like to finish this interesting mechanism and complete the other steps. The second step is currently underway, and showing promising results as well. We had, however, to build a new model based on the intermediary of the first step, in which we now have in the active site Cys161 bound to the acyl group and new malonyl moiety bound to PPT.

REFERENCES

1. Lehninger, A. L.; Nelson, D. L.; Cox, M. M., *Lehninger principles of biochemistry*. 5th ed.; W.H. Freeman: New York, 2008; p 1 v. (various pagings).
2. Mahler, H. In *Role of coenzyme A in fatty acid metabolism*, Federation proceedings, 1953; pp 694-702.
3. Lynen, F. In *Functional group of coenzyme A and its metabolic relations, especially in the fatty acid cycle*, Federation proceedings, 1953; pp 683-691.
4. Majerus, P. W.; Alberts, A. W.; Vagelos, P. R., The acyl carrier protein of fatty acid synthesis: purification, physical properties, and substrate binding site. *Proceedings of the National Academy of Sciences* **1964,** *51* (6), 1231-1238.
5. Hussain, M. M.; Strickland, D. K.; Bakillah, A., The mammalian low-density lipoprotein receptor family. *Annu Rev Nutr* **1999,** *19*, 141-72.
6. Willnow, T. E., The low-density lipoprotein receptor gene family: multiple roles in lipid metabolism. *J Mol Med (Berl)* **1999,** *77* (3), 306-15.
7. Lecerf, J. M.; de Lorgeril, M., Dietary cholesterol: from physiology to cardiovascular risk. *Br J Nutr* **2011,** *106* (1), 6-14.
8. Libby, P., Inflammation in atherosclerosis. *Nature* **2002,** *420* (6917), 868-74.
9. Parks, L. W., Metabolism of sterols in yeast. *CRC Crit Rev Microbiol* **1978,** *6* (4), 301-41.
10. Goldstein, J. L.; Brown, M. S., Regulation of the mevalonate pathway. *Nature* **1990,** *343* (6257), 425-30.
11. Chappell, J.; Wolf, F.; Proulx, J.; Cuellar, R.; Saunders, C., Is the Reaction Catalyzed by 3-Hydroxy-3-Methylglutaryl Coenzyme A Reductase a Rate-Limiting Step for Isoprenoid Biosynthesis in Plants? *Plant Physiol* **1995,** *109* (4), 1337-1343.
12. Holstein, S. A.; Hohl, R. J., Isoprenoids: remarkable diversity of form and function. *Lipids* **2004,** *39* (4), 293-309.
13. Bochar, D. A.; Stauffacher, C. V.; Rodwell, V. W., Sequence comparisons reveal two classes of 3-hydroxy-3-methylglutaryl coenzyme A reductase. *Mol Genet Metab* **1999,** *66* (2), 122-7.
14. Hedl, M.; Tabernero, L.; Stauffacher, C. V.; Rodwell, V. W., Class II 3-hydroxy-3-methylglutaryl coenzyme A reductases. *J Bacteriol* **2004,** *186* (7), 1927-32.
15. Beach, M. J.; Rodwell, V. W., Cloning, sequencing, and overexpression of mvaA, which encodes Pseudomonas mevalonii 3-hydroxy-3-methylglutaryl coenzyme A reductase. *J Bacteriol* **1989,** *171* (6), 2994-3001.
16. Istvan, E. S.; Palnitkar, M.; Buchanan, S. K.; Deisenhofer, J., Crystal structure of the catalytic portion of human HMG-CoA reductase: insights into regulation of activity and catalysis. *EMBO J* **2000,** *19* (5), 819-30.
17. Istvan, E. S.; Deisenhofer, J., The structure of the catalytic portion of human HMG-CoA reductase. *Biochim Biophys Acta* **2000,** *1529* (1-3), 9-18.
18. Brown, M. S.; Goldstein, J. L., A proteolytic pathway that controls the cholesterol content of membranes, cells, and blood. *Proc Natl Acad Sci U S A* **1999,** *96* (20), 11041-8.
19. Frimpong, K.; Rodwell, V. W., Catalysis by Syrian hamster 3-hydroxy-3-methylglutaryl-coenzyme A reductase. Proposed roles of histidine 865, glutamate 558, and aspartate 766. *J Biol Chem* **1994,** *269* (15), 11478-83.
20. Rajavashisth, T. B.; Taylor, A. K.; Andalibi, A.; Svenson, K. L.; Lusis, A. J., Identification of a zinc finger protein that binds to the sterol regulatory element. *Science* **1989,** *245* (4918), 640-3.
21. Osborne, T. F.; Gil, G.; Goldstein, J. L.; Brown, M. S., Operator constitutive mutation of 3-hydroxy-3-methylglutaryl coenzyme A reductase promoter abolishes protein binding to sterol regulatory element. *J Biol Chem* **1988,** *263* (7), 3380-7.

22. Tanaka, R. D.; Edwards, P. A.; Lan, S. F.; Fogelman, A. M., Regulation of 3-hydroxy-3-methylglutaryl coenzyme A reductase activity in avian myeloblasts. Mode of action of 25-hydroxycholesterol. *J Biol Chem* **1983,** *258* (21), 13331-9.
23. Faust, J. R.; Luskey, K. L.; Chin, D. J.; Goldstein, J. L.; Brown, M. S., Regulation of synthesis and degradation of 3-hydroxy-3-methylglutaryl-coenzyme A reductase by low density lipoprotein and 25-hydroxycholesterol in UT-1 cells. *Proc Natl Acad Sci U S A* **1982,** *79* (17), 5205-9.
24. Beg, Z. H.; Stonik, J. A.; Brewer, H. B., Jr., In vivo modulation of rat liver 3-hydroxy-3-methylglutaryl-coenzyme A reductase, reductase kinase, and reductase kinase kinase by mevalonolactone. *Proc Natl Acad Sci U S A* **1984,** *81* (23), 7293-7.
25. Horton, J. D.; Goldstein, J. L.; Brown, M. S., SREBPs: activators of the complete program of cholesterol and fatty acid synthesis in the liver. *J Clin Invest* **2002,** *109* (9), 1125-31.
26. Preiss, B., *Regulation of HMG-CoA reductase*. Academic Press: Orlando, 1985; p xii, 330 p.
27. Sever, N.; Yang, T.; Brown, M. S.; Goldstein, J. L.; DeBose-Boyd, R. A., Accelerated degradation of HMG CoA reductase mediated by binding of insig-1 to its sterol-sensing domain. *Mol Cell* **2003,** *11* (1), 25-33.
28. Sever, N.; Song, B. L.; Yabe, D.; Goldstein, J. L.; Brown, M. S.; DeBose-Boyd, R. A., Insig-dependent ubiquitination and degradation of mammalian 3-hydroxy-3-methylglutaryl-CoA reductase stimulated by sterols and geranylgeraniol. *J Biol Chem* **2003,** *278* (52), 52479-90.
29. Omkumar, R. V.; Darnay, B. G.; Rodwell, V. W., Modulation of Syrian hamster 3-hydroxy-3-methylglutaryl-CoA reductase activity by phosphorylation. Role of serine 871. *J Biol Chem* **1994,** *269* (9), 6810-4.
30. Panda, T.; Devi, V. A., Regulation and degradation of HMGCo-A reductase. *Appl Microbiol Biotechnol* **2004,** *66* (2), 143-52.
31. Omkumar, R. V.; Rodwell, V. W., Phosphorylation of Ser871 impairs the function of His865 of Syrian hamster 3-hydroxy-3-methylglutaryl-CoA reductase. *J Biol Chem* **1994,** *269* (24), 16862-6.
32. LaRosa, J. C., Low-density lipoprotein cholesterol reduction: the end is more important than the means. *Am J Cardiol* **2007,** *100* (2), 240-2.
33. Hegsted, D. M., Serum-cholesterol response to dietary cholesterol: a re-evaluation. *Am J Clin Nutr* **1986,** *44* (2), 299-305.
34. Endo, A.; Kuroda, M.; Tanzawa, K., Competitive inhibition of 3-hydroxy-3-methylglutaryl coenzyme A reductase by ML-236A and ML-236B fungal metabolites, having hypocholesterolemic activity. *FEBS Lett* **1976,** *72* (2), 323-6.
35. Endo, A.; Kuroda, M.; Tsujita, Y., ML-236A, ML-236B, and ML-236C, new inhibitors of cholesterogenesis produced by Penicillium citrinium. *J Antibiot (Tokyo)* **1976,** *29* (12), 1346-8.
36. Gotto, A. M., Jr., Results of recent large cholesterol-lowering trials and implications for clinical management. *Am J Cardiol* **1997,** *79* (12), 1663-6.
37. Brown, M. S.; Faust, J. R.; Goldstein, J. L.; Kaneko, I.; Endo, A., Induction of 3-hydroxy-3-methylglutaryl coenzyme A reductase activity in human fibroblasts incubated with compactin (ML-236B), a competitive inhibitor of the reductase. *J Biol Chem* **1978,** *253* (4), 1121-8.
38. Istvan, E. S.; Deisenhofer, J., Structural mechanism for statin inhibition of HMG-CoA reductase. *Science* **2001,** *292* (5519), 1160-4.
39. Alberts, A. W.; Chen, J.; Kuron, G.; Hunt, V.; Huff, J.; Hoffman, C.; Rothrock, J.; Lopez, M.; Joshua, H.; Harris, E.; Patchett, A.; Monaghan, R.; Currie, S.; Stapley, E.; Albers-Schonberg, G.; Hensens, O.; Hirshfield, J.; Hoogsteen, K.; Liesch, J.; Springer, J., Mevinolin: a highly potent competitive inhibitor of hydroxymethylglutaryl-coenzyme A reductase and a cholesterol-lowering agent. *Proc Natl Acad Sci U S A* **1980,** *77* (7), 3957-61.

40. Tobert, J. A., Lovastatin and beyond: the history of the HMG-CoA reductase inhibitors. *Nat Rev Drug Discov* **2003,** *2* (7), 517-26.
41. Singh, N.; Tamariz, J.; Chamorro, G.; Medina-Franco, J. L., Inhibitors of HMG-CoA Reductase: Current and Future Prospects. *Mini Rev Med Chem* **2009,** *9* (11), 1272-83.
42. Golomb, B. A.; Evans, M. A., Statin adverse effects : a review of the literature and evidence for a mitochondrial mechanism. *Am J Cardiovasc Drugs* **2008,** *8* (6), 373-418.
43. Skottheim, I. B.; Gedde-Dahl, A.; Hejazifar, S.; Hoel, K.; Asberg, A., Statin induced myotoxicity: the lactone forms are more potent than the acid forms in human skeletal muscle cells in vitro. *Eur J Pharm Sci* **2008,** *33* (4-5), 317-25.
44. Carbonell, T.; Freire, E., Binding thermodynamics of statins to HMG-CoA reductase. *Biochemistry* **2005,** *44* (35), 11741-8.
45. Wang, C. Y.; Liu, P. Y.; Liao, J. K., Pleiotropic effects of statin therapy: molecular mechanisms and clinical results. *Trends Mol Med* **2008,** *14* (1), 37-44.
46. Graham, D. J.; Staffa, J. A.; Shatin, D.; Andrade, S. E.; Schech, S. D.; La Grenade, L.; Gurwitz, J. H.; Chan, K. A.; Goodman, M. J.; Platt, R., Incidence of hospitalized rhabdomyolysis in patients treated with lipid-lowering drugs. *JAMA* **2004,** *292* (21), 2585-90.
47. Berman, H.; Henrick, K.; Nakamura, H., Announcing the worldwide Protein Data Bank. *Nat Struct Biol* **2003,** *10* (12), 980-980.
48. Dennington, R.; Keith, T.; Millam, J., GaussView, version 5. *Semichem Inc., Shawnee Mission, KS* **2009**.
49. Cramer, C. J., *Essentials of Computational Chemistry: Theories and Models*. Wiley: 2005.
50. Jensen, F., *Introduction to Computational Chemistry*. Wiley: 2006.
51. Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A., A 2nd Generation Force-Field for the Simulation of Proteins, Nucleic-Acids, and Organic-Molecules. *J Am Chem Soc* **1995,** *117* (19), 5179-5197.
52. Wang, J. M.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A., Development and testing of a general amber force field. *J Comput Chem* **2004,** *25* (9), 1157-1174.
53. Schlick, T., *Molecular Modeling and Simulation: An Interdisciplinary Guide: An Interdisciplinary Guide*. Springer: 2010.
54. Halgren, T. A.; Damm, W., Polarizable force fields. *Curr Opin Struc Biol* **2001,** *11* (2), 236-242.
55. Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L., Comparison of simple potential functions for simulating liquid water. *The Journal of chemical physics* **1983,** *79* (2), 926-935.
56. Bayly, C. I.; Cieplak, P.; Cornell, W. D.; Kollman, P. A., A Well-Behaved Electrostatic Potential Based Method Using Charge Restraints for Deriving Atomic Charges - the Resp Model. *J Phys Chem-Us* **1993,** *97* (40), 10269-10280.
57. Atkins, P. W. F., R. S., *Molecular Quantum Mechanics*. OUP Oxford: 2011.
58. Griffiths, D. J., *Introduction to quantum mechanics*. Pearson Prentice Hall: 2005.
59. Hayward, D. O., *Quantum Mechanics for Chemists*. Royal Society of Chemistry: 2002.
60. McQuarrie, D. A., *Quantum Chemistry*. University Science Books: 2008.
61. Binkley, J. S.; Pople, J. A., Moller-Plesset Theory for Atomic Ground-State Energies. *Int J Quantum Chem* **1975,** *9* (2), 229-236.
62. Shavitt, I., The history and evolution of configuration interaction. *Mol Phys* **1998,** *94* (1), 3-17.
63. Scuseria, G. E., The Open-Shell Restricted Hartree-Fock Singles and Doubles Coupled-Cluster Method Including Triple Excitations Ccsd (T) - Application to C3+. *Chem Phys Lett* **1991,** *176* (1), 27-35.

64. Dapprich, S.; Komaromi, I.; Byun, K. S.; Morokuma, K.; Frisch, M. J., A new ONIOM implementation in Gaussian98. Part I. The calculation of energies, gradients, vibrational frequencies and electric field derivatives. *J Mol Struc-Theochem* **1999,** *461*, 1-21.
65. Maseras, F.; Morokuma, K., Imomm - a New Integrated Ab-Initio Plus Molecular Mechanics Geometry Optimization Scheme of Equilibrium Structures and Transition-States. *J Comput Chem* **1995,** *16* (9), 1170-1179.
66. Svensson, M.; Humbel, S.; Froese, R. D. J.; Matsubara, T.; Sieber, S.; Morokuma, K., ONIOM: A multilayered integrated MO+MM method for geometry optimizations and single point energy predictions. A test for Diels-Alder reactions and Pt(P(t-Bu)(3))(2)+H-2 oxidative addition. *J Phys Chem-Us* **1996,** *100* (50), 19357-19363.
67. Vreven, T.; Byun, K. S.; Komaromi, I.; Dapprich, S.; Montgomery, J. A.; Morokuma, K.; Frisch, M. J., Combining quantum mechanics methods with molecular mechanics methods in ONIOM. *J Chem Theory Comput* **2006,** *2* (3), 815-826.
68. Cerqueira, N. M.; Gesto, D.; Oliveira, E. F.; Santos-Martins, D.; Brás, N. F.; Sousa, S. F.; Fernandes, P. A.; Ramos, M. J., Receptor-based virtual screening protocol for drug discovery. *Arch Biochem Biophys* **2015,** *582*, 56-67.
69. Van Drie, J. H., Computer-aided drug design: the next 20 years. *J Comput Aid Mol Des* **2007,** *21* (10-11), 591-601.
70. Jorgensen, W. L., The many roles of computation in drug discovery. *Science* **2004,** *303* (5665), 1813-1818.
71. Ke, Y. Y.; Coumar, M. S.; Shiao, H. Y.; Wang, W. C.; Chen, C. W.; Song, J. S.; Chen, C. H.; Lin, W. H.; Wu, S. H.; Hsu, J. T. A.; Chang, C. M.; Hsieh, H. P., Ligand efficiency based approach for efficient virtual screening of compound libraries. *European journal of medicinal chemistry* **2014,** *83*, 226-235.
72. Kalyaanamoorthy, S.; Chen, Y. P. P., Structure-based drug design to augment hit discovery. *Drug discovery today* **2011,** *16* (17-18), 831-839.
73. Tanrikulu, Y.; Proschak, E.; Werner, T.; Geppert, T.; Todoroff, N.; Klenner, A.; Kottke, T.; Sander, K.; Schneider, E.; Seifert, R.; Stark, H.; Clark, T.; Schneider, G., Homology Model Adjustment and Ligand Screening with a Pseudoreceptor of the Human Histamine H-4 Receptor. *Chemmedchem* **2009,** *4* (5), 820-827.
74. Park, H.; Bahn, Y. J.; Ryu, S. E., Structure-based de novo design and biochemical evaluation of novel Cdc25 phosphatase inhibitors. *Bioorg Med Chem Lett* **2009,** *19* (15), 4330-4334.
75. Budzik, B.; Garzya, V.; Shi, D. C.; Walker, G.; Woolley-Roberts, M.; Pardoe, J.; Lucas, A.; Tehan, B.; Rivero, R. A.; Langmead, C. J.; Watson, J.; Wu, Z. N.; Forbes, I. T.; Jin, J. A., Novel N-Substituted Benzimidazolones as Potent, Selective, CNS-Penetrant, and Orally Active M-1 mAChR Agonists. *ACS medicinal chemistry letters* **2010,** *1* (6), 244-248.
76. Levitt, D. G.; Banaszak, L. J., Pocket - a Computer-Graphics Method for Identifying and Displaying Protein Cavities and Their Surrounding Amino-Acids. *J Mol Graphics* **1992,** *10* (4), 229-234.
77. Hendlich, M.; Rippmann, F.; Barnickel, G., LIGSITE: Automatic and efficient detection of potential small molecule-binding sites in proteins. *J Mol Graph Model* **1997,** *15* (6), 359-+.
78. Laskowski, R. A., SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. *Journal of molecular graphics* **1995,** *13* (5), 323-30, 307-8.
79. Desjarlais, R. L.; Sheridan, R. P.; Seibel, G. L.; Dixon, J. S.; Kuntz, I. D.; Venkataraghavan, R., Using Shape Complementarity as an Initial Screen in Designing Ligands for a Receptor-Binding Site of Known 3-Dimensional Structure. *J Med Chem* **1988,** *31* (4), 722-729.
80. Le Guilloux, V.; Schmidtke, P.; Tuffery, P., Fpocket: an open source platform for ligand pocket detection. *BMC bioinformatics* **2009,** *10*, 168.

81. Laurie, A. T.; Jackson, R. M., Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites. *Bioinformatics* **2005,** *21* (9), 1908-16.
82. Reynolds, C. A.; Wade, R. C.; Goodford, P. J., Identifying Targets for Bioreductive Agents - Using Grid to Predict Selective Binding Regions of Proteins. *J Mol Graphics* **1989,** *7* (2), 103-&.
83. Wade, R. C.; Goodford, P. J., Further Development of Hydrogen-Bond Functions for Use in Determining Energetically Favorable Binding-Sites on Molecules of Known Structure .2. Ligand Probe Groups with the Ability to Form More Than 2 Hydrogen-Bonds. *J Med Chem* **1993,** *36* (1), 148-156.
84. Weisel, M.; Proschak, E.; Schneider, G., PocketPicker: analysis of ligand binding-sites with shape descriptors. *Chemistry Central journal* **2007,** *1*, 7.
85. Henrich, S.; Salo-Ahen, O. M.; Huang, B.; Rippmann, F. F.; Cruciani, G.; Wade, R. C., Computational approaches to identifying and characterizing protein binding sites for ligand design. *J Mol Recognit* **2010,** *23* (2), 209-19.
86. Kortvelyesi, T.; Dennis, S.; Silberstein, M.; Brown, L., 3rd; Vajda, S., Algorithms for computational solvent mapping of proteins. *Proteins* **2003,** *51* (3), 340-51.
87. Schmidtke, P.; Barril, X., Understanding and Predicting Druggability. A High-Throughput Method for Detection of Drug Binding Sites. *J Med Chem* **2010,** *53* (15), 5858-5867.
88. Kortagere, S.; Ekins, S., Troubleshooting computational methods in drug discovery. *J Pharmacol Tox Met* **2010,** *61* (2), 67-75.
89. Nisius, B.; Sha, F.; Gohlke, H., Structure-based computational analysis of protein binding sites for function and druggability prediction. *J Biotechnol* **2012,** *159* (3), 123-134.
90. Sastry, G. M.; Adzhigirey, M.; Day, T.; Annabhimoju, R.; Sherman, W., Protein and ligand preparation: parameters, protocols, and influence on virtual screening enrichments. *J Comput Aided Mol Des* **2013,** *27* (3), 221-34.
91. Roberts, B. C.; Mancera, R. L., Ligand-protein docking with water molecules. *J Chem Inf Model* **2008,** *48* (2), 397-408.
92. Beuming, T.; Che, Y.; Abel, R.; Kim, B.; Shanmugasundaram, V.; Sherman, W., Thermodynamic analysis of water molecules at the surface of proteins and applications to binding site prediction and characterization. *Proteins* **2012,** *80* (3), 871-883.
93. Khandelwal, A.; Lukacova, V.; Comez, D.; Kroll, D. M.; Raha, S.; Balaz, S., A combination of docking, QM/MM methods, and MD simulation for binding affinity estimation of metalloprotein ligands. *J Med Chem* **2005,** *48* (17), 5437-47.
94. Strynadka, N. C.; Eisenstein, M.; Katchalski-Katzir, E.; Shoichet, B. K.; Kuntz, I. D.; Abagyan, R.; Totrov, M.; Janin, J.; Cherfils, J.; Zimmerman, F.; Olson, A.; Duncan, B.; Rao, M.; Jackson, R.; Sternberg, M.; James, M. N., Molecular docking programs successfully predict the binding of a beta-lactamase inhibitory protein to TEM-1 beta-lactamase. *Nature structural biology* **1996,** *3* (3), 233-9.
95. Williams, A. J., Public chemical compound databases. *Current opinion in drug discovery & development* **2008,** *11* (3), 393-404.
96. Irwin, J. J.; Sterling, T.; Mysinger, M. M.; Bolstad, E. S.; Coleman, R. G., ZINC: A Free Tool to Discover Chemistry for Biology. *J Chem Inf Model* **2012,** *52* (7), 1757-1768.
97. Chen, J.; Swamidass, S. J.; Bruand, J.; Baldi, P., ChemDB: a public database of small molecules and related chemoinformatics resources. *Bioinformatics* **2005,** *21* (22), 4133-4139.
98. Bolton, E. E.; Wang, Y.; Thiessen, P. A.; Bryant, S. H., PubChem: integrated platform of small molecules and biological activities. *Annual reports in computational chemistry* **2008,** *4*, 217-241.
99. Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J., Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv Drug Deliver Rev* **1997,** *23* (1-3), 3-25.
100. Ghose, R. P.; Hall, P. M.; Bravo, E. L., Medical management of aldosterone-producing adenomas. *Ann Intern Med* **1999,** *131* (2), 105-+.

101. Veber, D. F.; Johnson, S. R.; Cheng, H. Y.; Smith, B. R.; Ward, K. W.; Kopple, K. D., Molecular properties that influence the oral bioavailability of drug candidates. *J Med Chem* **2002,** *45* (12), 2615-2623.
102. Kuntz, I. D.; Blaney, J. M.; Oatley, S. J.; Langridge, R.; Ferrin, T. E., A GEOMETRIC APPROACH TO MACROMOLECULE-LIGAND INTERACTIONS. *Journal of Molecular Biology* **1982,** *161* (2), 269-288.
103. Ritchie, D. W., Recent progress and future directions in protein-protein docking. *Current Protein & Peptide Science* **2008,** *9* (1), 1-15.
104. Hashmi, I.; Shehu, A., HopDock: a probabilistic search algorithm for decoy sampling in protein-protein docking. *Proteome Science* **2013,** *11*.
105. Chen, R.; Li, L.; Weng, Z. P., ZDOCK: An initial-stage protein-docking algorithm. *Proteins-Structure Function and Genetics* **2003,** *52* (1), 80-87.
106. Gabb, H. A.; Jackson, R. M.; Sternberg, M. J. E., Modelling protein docking using shape complementarity, electrostatics and biochemical information. *Journal of Molecular Biology* **1997,** *272* (1), 106-120.
107. Pang, Y. P.; Kozikowski, A. P., PREDICTION OF THE BINDING-SITE OF 1-BENZYL-4- (5,6-DIMETHOXY-1-INDANON-2-YL)METHYL PIPERIDINE IN ACETYLCHOLINESTERASE BY DOCKING STUDIES WITH THE SYSDOC PROGRAM. *Journal of Computer-Aided Molecular Design* **1994,** *8* (6), 683-693.
108. Perola, E.; Xu, K.; Kollmeyer, T. M.; Kaufmann, S. H.; Prendergast, F. G.; Pang, Y. P., Successful virtual screening of a chemical database for farnesyltransferase inhibitor leads. *J Med Chem* **2000,** *43* (3), 401-408.
109. Ewing, T. J. A.; Makino, S.; Skillman, A. G.; Kuntz, I. D., DOCK 4.0: Search strategies for automated molecular docking of flexible molecule databases. *Journal of Computer-Aided Molecular Design* **2001,** *15* (5), 411-428.
110. Sauton, N.; Lagorce, D.; Villoutreix, B. O.; Miteva, M. A., MS-DOCK: Accurate multiple conformation generator and rigid docking protocol for multi-step virtual ligand screening. *Bmc Bioinformatics* **2008,** *9*.
111. Venkatachalam, C. M.; Jiang, X.; Oldfield, T.; Waldman, M., LigandFit: a novel method for the shape-directed rapid docking of ligands to protein active sites. *Journal of Molecular Graphics & Modelling* **2003,** *21* (4), 289-307.
112. Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S., Glide: A new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *Journal of Medicinal Chemistry* **2004,** *47* (7), 1739-1749.
113. Amaro, R. E.; Baron, R.; McCammon, J. A., An improved relaxed complex scheme for receptor flexibility in computer-aided drug design. *Journal of Computer-Aided Molecular Design* **2008,** *22* (9), 693-705.
114. Bolstad, E. S. D.; Anderson, A. C., In pursuit of virtual lead optimization: Pruning ensembles of receptor structures for increased efficiency and accuracy during docking. *Proteins-Structure Function and Bioinformatics* **2009,** *75* (1), 62-74.
115. Cavasotto, C. N.; Singh, N., Docking and high throughput docking: Successes and the challenge of protein flexibility. *Current Computer-Aided Drug Design* **2008,** *4* (3), 221-234.
116. Bohm, H. J., THE COMPUTER-PROGRAM LUDI - A NEW METHOD FOR THE DENOVO DESIGN OF ENZYME-INHIBITORS. *Journal of Computer-Aided Molecular Design* **1992,** *6* (1), 61-78.
117. Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G., A fast flexible docking method using an incremental construction algorithm. *J Mol Biol* **1996,** *261* (3), 470-489.
118. Mizutani, M. Y.; Tomioka, N.; Itai, A., RATIONAL AUTOMATIC SEARCH METHOD FOR STABLE DOCKING MODELS OF PROTEIN AND LIGAND. *Journal of Molecular Biology* **1994,** *243* (2), 310-326.
119. Welch, W.; Ruppert, J.; Jain, A. N., Hammerhead: Fast, fully automated docking of flexible ligands to protein binding sites. *Chemistry & Biology* **1996,** *3* (6), 449-462.

120. Jain, A. N., Surflex-Dock 2.1: Robust performance from ligand energetic modeling, ring flexibility, and knowledge-based search. *Journal of Computer-Aided Molecular Design* **2007,** *21* (5), 281-306.
121. Jain, A. N., Surflex: Fully automatic flexible molecular docking using a molecular similarity-based search engine. *Journal of Medicinal Chemistry* **2003,** *46* (4), 499-511.
122. Zsoldos, Z.; Reid, D.; Simon, A.; Sadjad, S. B.; Johnson, A. P., eHiTS: A new fast, exhaustive flexible ligand docking system. *Journal of Molecular Graphics & Modelling* **2007,** *26* (1), 198-212.
123. Miller, M. D.; Kearsley, S. K.; Underwood, D. J.; Sheridan, R. P., FLOG - A SYSTEM TO SELECT QUASI-FLEXIBLE LIGANDS COMPLEMENTARY TO A RECEPTOR OF KNOWN 3-DIMENSIONAL STRUCTURE. *Journal of Computer-Aided Molecular Design* **1994,** *8* (2), 153-174.
124. Trosset, J. Y.; Scheraga, H. A., PRODOCK: Software package for protein modeling and docking. *Journal of Computational Chemistry* **1999,** *20* (4), 412-427.
125. Abagyan, R.; Totrov, M.; Kuznetsov, D., ICM - A NEW METHOD FOR PROTEIN MODELING AND DESIGN - APPLICATIONS TO DOCKING AND STRUCTURE PREDICTION FROM THE DISTORTED NATIVE CONFORMATION. *Journal of Computational Chemistry* **1994,** *15* (5), 488-506.
126. Liu, M.; Wang, S. M., MCDOCK: A Monte Carlo simulation approach to the molecular docking problem. *Journal of Computer-Aided Molecular Design* **1999,** *13* (5), 435-451.
127. Hart, T. N.; Read, R. J., A MULTIPLE-START MONTE-CARLO DOCKING METHOD. *Proteins-Structure Function and Genetics* **1992,** *13* (3), 206-222.
128. McMartin, C.; Bohacek, R. S., QXP: Powerful, rapid computer algorithms for structure-based drug design. *Journal of Computer-Aided Molecular Design* **1997,** *11* (4), 333-344.
129. Jones, G.; Willett, P.; Glen, R. C., MOLECULAR RECOGNITION OF RECEPTOR-SITES USING A GENETIC ALGORITHM WITH A DESCRIPTION OF DESOLVATION. *Journal of Molecular Biology* **1995,** *245* (1), 43-53.
130. Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R., Development and validation of a genetic algorithm for flexible docking. *Journal of Molecular Biology* **1997,** *267* (3), 727-748.
131. Morris, G. M.; Goodsell, D. S.; Halliday, R. S.; Huey, R.; Hart, W. E.; Belew, R. K.; Olson, A. J., Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *Journal of Computational Chemistry* **1998,** *19* (14), 1639-1662.
132. Clark, K. P.; Ajay, FLEXIBLE LIGAND DOCKING WITHOUT PARAMETER ADJUSTMENT ACROSS 4 LIGAND-RECEPTOR COMPLEXES. *Journal of Computational Chemistry* **1995,** *16* (10), 1210-1226.
133. Taylor, J. S.; Burnett, R. M., DARWIN: A program for docking flexible molecules. *Proteins-Structure Function and Genetics* **2000,** *41* (2), 173-191.
134. Murray, C. W.; Baxter, C. A.; Frenkel, A. D., The sensitivity of the results of molecular docking to induced fit effects: Application to thrombin, thermolysin and neuraminidase. *Journal of Computer-Aided Molecular Design* **1999,** *13* (6), 547-562.
135. Baxter, C. A.; Murray, C. W.; Clark, D. E.; Westhead, D. R.; Eldridge, M. D., Flexible docking using Tabu search and an empirical estimate of binding affinity. *Proteins-Structure Function and Genetics* **1998,** *33* (3), 367-382.
136. Cerqueira, N. M.; Bras, N. F.; Fernandes, P. A.; Ramos, M. J., MADAMM: a multistaged docking with an automated molecular modeling protocol. *Proteins* **2009,** *74* (1), 192-206.
137. Apostolakis, J.; Pluckthun, A.; Caflisch, A., Docking small ligands in flexible binding sites. *Journal of Computational Chemistry* **1998,** *19* (1), 21-37.
138. Pak, Y. S.; Wang, S. M., Application of a molecular dynamics simulation method with a generalized effective potential to the flexible molecular docking problems. *Journal of Physical Chemistry B* **2000,** *104* (2), 354-359.

139. Schnecke, V.; Swanson, C. A.; Getzoff, E. D.; Tainer, J. A.; Kuhn, L. A., Screening a peptidyl database for potential ligands to proteins with side-chain flexibility. *Proteins-Structure Function and Genetics* **1998,** *33* (1), 74-87.
140. Cerqueira, N. M. F. S. A.; Bras, N. F.; Fernandes, P. A.; Ramos, M. J., MADAMM: A multistaged docking with an automated molecular modeling protocol. *Proteins-Structure Function and Bioinformatics* **2009,** *74* (1), 192-206.
141. Knegtel, R. M. A.; Kuntz, I. D.; Oshiro, C. M., Molecular docking to ensembles of protein structures. *Journal of Molecular Biology* **1997,** *266* (2), 424-440.
142. Carlson, H. A., Protein flexibility is an important component of structure-based drug discovery. *Current Pharmaceutical Design* **2002,** *8* (17), 1571-1578.
143. Claussen, H.; Buning, C.; Rarey, M.; Lengauer, T., FlexE: Efficient molecular docking considering protein structure variations. *Journal of Molecular Biology* **2001,** *308* (2), 377-395.
144. Osterberg, F.; Morris, G. M.; Sanner, M. F.; Olson, A. J.; Goodsell, D. S., Automated docking to multiple target structures: Incorporation of protein mobility and structural water heterogeneity in AutoDock. *Proteins-Structure Function and Genetics* **2002,** *46* (1), 34-40.
145. Morris, G. M.; Goodsell, D. S.; Huey, R.; Olson, A. J., Distributed automated docking of flexible ligands to proteins: parallel applications of AutoDock 2.4. *J Comput Aided Mol Des* **1996,** *10* (4), 293-304.
146. Viegas, A.; Bras, N. F.; Cerqueira, N. M. F. S. A.; Fernandes, P. A.; Prates, J. A. M.; Fontes, C. M. G. A.; Bruix, M.; Romao, M. J.; Carvalho, A. L.; Ramos, M. J.; Macedo, A. L.; Cabrita, E. J., Molecular determinants of ligand specificity in family 11 carbohydrate binding modules - an NMR, X-ray crystallography and computational chemistry approach. *Febs J* **2008,** *275* (10), 2524-2535.
147. Bras, N. F.; Cerqueira, N. M. F. S. A.; Fernandes, P. A.; Ramos, M. J., Carbohydrate-binding modules from family 11: Understanding the binding mode of polysaccharides. *Int J Quantum Chem* **2008,** *108* (11), 2030-2040.
148. Yuriev, E.; Ramsland, P. A., Latest developments in molecular docking: 2010-2011 in review. *J Mol Recognit* **2013,** *26* (5), 215-239.
149. Kramer, B.; Rarey, M.; Lengauer, T., Evaluation of the FLEXX incremental construction algorithm for protein-ligand docking. *Proteins* **1999,** *37* (2), 228-41.
150. Makino, S.; Kuntz, I. D., Automated flexible ligand docking method and its application for database search. *J Comput Chem* **1997,** *18* (14), 1812-1825.
151. Morris, G. M.; Huey, R.; Olson, A. J., Using AutoDock for ligand-receptor docking. *Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.]* **2008,** *Chapter 8*, Unit 8 14.
152. Weiner, S. J.; Kollman, P. A.; Nguyen, D. T.; Case, D. A., An All Atom Force-Field for Simulations of Proteins and Nucleic-Acids. *J Comput Chem* **1986,** *7* (2), 230-252.
153. Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G., A fast flexible docking method using an incremental construction algorithm. *J Mol Biol* **1996,** *261* (3), 470-89.
154. Gehlhaar, D. K.; Verkhivker, G. M.; Rejto, P. A.; Sherman, C. J.; Fogel, D. B.; Fogel, L. J.; Freer, S. T., Molecular Recognition of the Inhibitor Ag-1343 by Hiv-1 Protease - Conformationally Flexible Docking by Evolutionary Programming. *Chem Biol* **1995,** *2* (5), 317-324.
155. Eldridge, M. D.; Murray, C. W.; Auton, T. R.; Paolini, G. V.; Mee, R. P., Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *J Comput Aided Mol Des* **1997,** *11* (5), 425-45.
156. Murray, C. W.; Auton, T. R.; Eldridge, M. D., Empirical scoring functions. II. The testing of an empirical scoring function for the prediction of ligand-receptor binding affinities and the use of Bayesian regression to improve the quality of the model. *J Comput Aided Mol Des* **1998,** *12* (5), 503-19.
157. Friesner, R. A.; Murphy, R. B.; Repasky, M. P.; Frye, L. L.; Greenwood, J. R.; Halgren, T. A.; Sanschagrin, P. C.; Mainz, D. T., Extra precision glide: Docking and

scoring incorporating a model of hydrophobic enclosure for protein-ligand complexes. *J Med Chem* **2006,** *49* (21), 6177-6196.
158. Wang, R. X.; Liu, L.; Lai, L. H.; Tang, Y. Q., SCORE: A new empirical method for estimating the binding affinity of a protein-ligand complex. *J Mol Model* **1998,** *4* (12), 379-394.
159. Rognan, D.; Lauemoller, S. L.; Holm, A.; Buus, S.; Tschinke, V., Predicting binding affinities of protein ligands from three-dimensional models: application to peptide binding to class I major histocompatibility proteins. *J Med Chem* **1999,** *42* (22), 4650-8.
160. Wang, R. X.; Lai, L. H.; Wang, S. M., Further development and validation of empirical scoring functions for structure-based binding affinity prediction. *J Comput Aid Mol Des* **2002,** *16* (1), 11-26.
161. Muegge, I., PMF scoring revisited. *J Med Chem* **2006,** *49* (20), 5895-5902.
162. Muegge, I.; Martin, Y. C., A general and fast scoring function for protein-ligand interactions: A simplified potential approach. *J Med Chem* **1999,** *42* (5), 791-804.
163. Muegge, I., Effect of ligand volume correction on PMF scoring. *J Comput Chem* **2001,** *22* (4), 418-425.
164. Gohlke, H.; Hendlich, M.; Klebe, G., Knowledge-based scoring function to predict protein-ligand interactions. *J Mol Biol* **2000,** *295* (2), 337-356.
165. Velec, H. F. G.; Gohlke, H.; Klebe, G., DrugScore(CSD)-knowledge-based scoring function derived from small molecule crystal data with superior recognition rate of near-native ligand poses and better affinity prediction. *J Med Chem* **2005,** *48* (20), 6296-6303.
166. Ishchenko, A. V.; Shakhnovich, E. I., SMall molecule growth 2001 (SMoG2001): An improved knowledge-based scoring function for protein-ligand interactions. *J Med Chem* **2002,** *45* (13), 2770-2780.
167. Trott, O.; Olson, A. J., AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J Comput Chem* **2010,** *31* (2), 455-61.
168. Hawkins, P. C. D.; Skillman, A. G.; Nicholls, A., Comparison of shape-matching and docking as virtual screening tools. *J Med Chem* **2007,** *50* (1), 74-82.
169. Ghosh, S.; Nie, A. H.; An, J.; Huang, Z. W., Structure-based virtual screening of chemical libraries for drug discovery. *Curr Opin Chem Biol* **2006,** *10* (3), 194-202.
170. Kitchen, D. B.; Decornez, H.; Furr, J. R.; Bajorath, J., Docking and scoring in virtual screening for drug discovery: Methods and applications. *Nat Rev Drug Discov* **2004,** *3* (11), 935-949.
171. Irwin, J. J.; Huang, N.; Shoichet, B., COMP 148-Benchmarking sets for molecular docking. *Abstr Pap Am Chem S* **2007,** *234*.
172. Huang, N.; Shoichet, B. K.; Irwin, J. J., Benchmarking sets for molecular docking. *J Med Chem* **2006,** *49* (23), 6789-6801.
173. Kirchmair, J.; Markt, P.; Distinto, S.; Wolber, G.; Langer, T., Evaluation of the performance of 3D virtual screening protocols: RMSD comparisons, enrichment assessments, and decoy selection - What can we learn from earlier mistakes? *J Comput Aid Mol Des* **2008,** *22* (3-4), 213-228.
174. Schneider, G., Virtual screening: an endless staircase? *Nat Rev Drug Discov* **2010,** *9* (4), 273-276.
175. Good, A. C.; Oprea, T. I., Optimization of CAMD techniques 3. Virtual screening enrichment studies: a help or hindrance in tool selection? *J Comput Aid Mol Des* **2008,** *22* (3-4), 169-178.
176. Ferrara, P.; Gohlke, H.; Price, D. J.; Klebe, G.; Brooks, C. L., 3rd, Assessing scoring functions for protein-ligand interactions. *J Med Chem* **2004,** *47* (12), 3032-47.
177. Cosconati, S.; Forli, S.; Perryman, A. L.; Harris, R.; Goodsell, D. S.; Olson, A. J., Virtual screening with AutoDock: theory and practice. *Expert opinion on drug discovery* **2010,** *5* (6), 597-607.

178. Charifson, P. S.; Corkery, J. J.; Murcko, M. A.; Walters, W. P., Consensus scoring: A method for obtaining improved hit rates from docking databases of three-dimensional structures into proteins. *J Med Chem* **1999,** *42* (25), 5100-5109.
179. Paul, N.; Rognan, D., ConsDock: A new program for the consensus analysis of protein-ligand interactions. *Proteins-Structure Function and Genetics* **2002,** *47* (4), 521-533.
180. Kapetanovic, I. M., Computer-aided drug discovery and development (CADDD): In silico-chemico-biological approach. *Chem-Biol Interact* **2008,** *171* (2), 165-176.
181. Cerqueira, N. M. F. S. A.; Oliveira, E. F.; Gesto, D. S.; Santos-Martins, D.; Moreira, C.; Moorthy, H. N.; Ramos, M. J.; Fernandes, P. A., Cholesterol Biosynthesis: A Mechanistic Overview. *Biochemistry* **2016,** *55* (39), 5483-5506.
182. Gordon, D. J.; Probstfield, J. L.; Garrison, R. J.; Neaton, J. D.; Castelli, W. P.; Knoke, J. D.; Jacobs, D. R., Jr.; Bangdiwala, S.; Tyroler, H. A., High-density lipoprotein cholesterol and cardiovascular disease. Four prospective American studies. *Circulation* **1989,** *79* (1), 8-15.
183. Briel, M.; Ferreira-Gonzalez, I.; You, J. J.; Karanicolas, P. J.; Akl, E. A.; Wu, P.; Blechacz, B.; Bassler, D.; Wei, X. G.; Sharman, A.; Whitt, I.; da Silva, S. A.; Khalid, Z.; Nordmann, A. J.; Zhou, Q.; Walter, S. D.; Vale, N.; Bhatnagar, N.; O'Regan, C.; Mills, E. J.; Bucher, H. C.; Montori, V. M.; Guyatt, G. H., Association between change in high density lipoprotein cholesterol and cardiovascular disease morbidity and mortality: systematic review and meta-regression analysis. *Brit Med J* **2009,** *338*.
184. Aarden, E.; Van Hoyweghen, I.; Horstman, K., The paradox of public health genomics: Definition and diagnosis of familial hypercholesterolaemia in three European countries. *Scand J Public Healt* **2011,** *39* (6), 634-639.
185. Newby, L. K.; Kandzari, D.; Roe, M. T.; Mulgund, J.; Bhatt, D. L.; DeLong, E.; Ohman, E. M.; Gibler, W. B.; Peterson, E. D., Examining the hypercholesterolemia paradox in acute coronary syndromes: Results from CRUSADE. *Circulation* **2005,** *111* (20), E329-E329.
186. Hussain, M. M.; Strickland, D. K.; Bakillah, A., The mammalian low-density lipoprotein receptor family. *Annu Rev Nutr* **1999,** *19*, 141-172.
187. Thompson, S.; Mayerl, F.; Peoples, O. P.; Masamune, S.; Sinskey, A. J.; Walsh, C. T., Mechanistic studies on beta-ketoacyl thiolase from Zoogloea ramigera: identification of the active-site nucleophile as Cys89, its mutation to Ser89, and kinetic and thermodynamic characterization of wild-type and mutant enzymes. *Biochemistry* **1989,** *28* (14), 5735-42.
188. Modis, Y.; Wierenga, R. K., Crystallographic analysis of the reaction pathway of Zoogloea ramigera biosynthetic thiolase. *J Mol Biol* **2000,** *297* (5), 1171-1182.
189. Antonenkov, V. D.; Van Veldhoven, P. P.; Waelkens, E.; Mannaerts, G. P., Substrate specificities of 3-oxoacyl-CoA thiolase A and sterol carrier protein 2/3-oxoacyl-CoA thiolase purified from normal rat liver peroxisomes. Sterol carrier protein 2/3-oxoacyl-CoA thiolase is involved in the metabolism of 2-methyl-branched fatty acids and bile acid intermediates. *J Biol Chem* **1997,** *272* (41), 26023-31.
190. Kovacs, W. J.; Tape, K. N.; Shackelford, J. E.; Duan, X. Y.; Kasumov, T.; Kelleher, J. K.; Brunengraber, H.; Krisans, S. K., Localization of the pre-squalene segment of the isoprenoid biosynthetic pathway in mammalian peroxisomes. *Histochem Cell Biol* **2007,** *127* (3), 273-290.
191. Tyni, T.; Palotie, A.; Viinikka, L.; Valanne, L.; Salo, M. K.; von Dobeln, U.; Jackson, S.; Wanders, R.; Venizelos, N.; Pihko, H., Long-chain 3-hydroxyacyl-coenzyme A dehydrogenase deficiency with the G1528C mutation: clinical presentation of thirteen patients. *The Journal of pediatrics* **1997,** *130* (1), 67-76.
192. Kamijo, T.; Indo, Y.; Souri, M.; Aoyama, T.; Hara, T.; Yamamoto, S.; Ushikubo, S.; Rinaldo, P.; Matsuda, I.; Komiyama, A.; Hashimoto, T., Medium chain 3-ketoacyl-coenzyme A thiolase deficiency: a new disorder of mitochondrial fatty acid beta-oxidation. *Pediatric research* **1997,** *42* (5), 569-76.

193. Hartlage, P.; Eller, G.; Carter, L.; Roesel, A.; Hommes, F., Mitochondrial acetoacetyl-CoA thiolase deficiency. *Biochemical medicine and metabolic biology* **1986,** *36* (2), 198-206.
194. Fukao, T.; Nakamura, H.; Song, X. Q.; Nakamura, K.; Orii, K. E.; Kohno, Y.; Kano, M.; Yamaguchi, S.; Hashimoto, T.; Orii, T.; Kondo, N., Characterization of N93S, I312T, and A333P missense mutations in two Japanese families with mitochondrial acetoacetyl-CoA thiolase deficiency. *Hum Mutat* **1998,** *12* (4), 245-254.
195. Harijan, R. K.; Kiema, T. R.; Karjalainen, M. P.; Janardan, N.; Murthy, M. R.; Weiss, M. S.; Michels, P. A.; Wierenga, R. K., Crystal structures of SCP2-thiolases of Trypanosomatidae, human pathogens causing widespread tropical diseases: the importance for catalysis of the cysteine of the unique HDCF loop. *The Biochemical journal* **2013,** *455* (1), 119-30.
196. Arnesen, T.; Thompson, P. R.; Varhaug, J. E.; Lillehaug, J. R., The Protein Acetyltransferase ARD1: A Novel Cancer Drug Target? *Curr Cancer Drug Tar* **2008,** *8* (7), 545-553.
197. Modis, Y.; Wierenga, R. K., A biosynthetic thiolase in complex with a reaction intermediate: the crystal structure provides new insights into the catalytic mechanism. *Structure* **1999,** *7* (10), 1279-90.
198. Merilainen, G.; Poikela, V.; Kursula, P.; Wierenga, R. K., The thiolase reaction mechanism: the importance of Asn316 and His348 for stabilizing the enolate intermediate of the Claisen condensation. *Biochemistry* **2009,** *48* (46), 11011-25.
199. Bahnson, B. J., An atomic-resolution mechanism of 3-hydroxy-3-methylglutaryl-CoA synthase. *P Natl Acad Sci USA* **2004,** *101* (47), 16399-16400.
200. Hegardt, F. G., Mitochondrial 3-hydroxy-3-methylglutaryl-CoA synthase: a control enzyme in ketogenesis. *Biochemical Journal* **1999,** *338*, 569-582.
201. Olivier, L. M.; Krisans, S. K., Peroxisomal protein targeting and identification of peroxisomal targeting signals in cholesterol biosynthetic enzymes. *Bba-Mol Cell Biol L* **2000,** *1529* (1-3), 89-102.
202. Olivier, L. M.; Kovacs, W.; Masuda, K.; Keller, G. A.; Krisans, S. K., Identification of peroxisomal targeting signals in cholesterol biosynthetic enzymes: AA-CoA thiolase, HMG-CoA synthase, MPPD, and FPP synthase. *J Lipid Res* **2000,** *41* (12), 1921-1935.
203. Sutherlin, A.; Hedl, M.; Sanchez-Neri, B.; Burgner, J. W.; Stauffacher, C. V.; Rodwell, V. W., Enterococcus faecalis 3-hydroxy-3-methylglutaryl coenzyme A synthase, an enzyme of isopentenyl diphosphate biosynthesis. *Journal of Bacteriology* **2002,** *184* (15), 4065-4070.
204. Shafqat, N.; Turnbull, A.; Zschocke, J.; Oppermann, U.; Yue, W. W., Crystal Structures of Human HMG-CoA Synthase Isoforms Provide Insights into Inherited Ketogenesis Disorders and Inhibitor Design. *J Mol Biol* **2010,** *398* (4), 497-506.
205. Misra, I.; Narasimhan, C.; Miziorko, H. M., Avian 3-Hydroxy-3-Methylglutaryl-Coa Synthase - Characterization of a Recombinant Cholesterogenic Isozyme and Demonstration of the Requirement for a Sulfhydryl Functionality in Formation of the Acetyl-Enzyme Reaction Intermediate. *Journal of Biological Chemistry* **1993,** *268* (16), 12129-12135.
206. Chun, K. Y.; Vinarov, D. A.; Miziorko, H. M., 3-Hydroxy-3-methylglutaryl-CoA synthase: participation of invariant acidic residues in formation of the acetyl-S-enzyme reaction intermediate. *Biochemistry* **2000,** *39* (47), 14670-81.
207. Miziorko, H. M.; Lane, M. D., 3-Hydroxy-3-methylgutaryl-CoA synthase. Participation of acetyl-S-enzyme and enzyme-S-hydroxymethylgutaryl-SCoA intermediates in the reaction. *J Biol Chem* **1977,** *252* (4), 1414-20.
208. Theisen, M. J.; Misra, I.; Saadat, D.; Campobasso, N.; Miziorko, H. M.; Harrison, D. H., 3-hydroxy-3-methylglutaryl-CoA synthase intermediate complex observed in "real-time". *Proc Natl Acad Sci U S A* **2004,** *101* (47), 16442-7.
209. Miziorko, H. M.; Clinkenbeard, K. D.; Reed, W. D.; Lane, M. D., 3-Hydroxy-3-methylglutaryl coenzyme A synthase. Evidence for an acetyl-S-enzyme intermediate and

identification of a cysteinyl sulfhydryl as the site of acetylation. *J Biol Chem* **1975,** *250* (15), 5768-73.
210. Bahnson, B. J., An atomic-resolution mechanism of 3-hydroxy-3-methylglutaryl-CoA synthase. *Proceedings of the National Academy of Sciences of the United States of America* **2004,** *101* (47), 16399-400.
211. Pojer, F.; Ferrer, J. L.; Richard, S. B.; Nagegowda, D. A.; Chye, M. L.; Bach, T. J.; Noel, J. P., Structural basis for the design of potent and species-specific inhibitors of 3-hydroxy-3-methylglutaryl CoA synthases. *P Natl Acad Sci USA* **2006,** *103* (31), 11491-11496.
212. Wilding, E. I.; Brown, J. R.; Bryant, A. P.; Chalker, A. F.; Holmes, D. J.; Ingraham, K. A.; Iordanescu, S.; Chi, Y. S.; Rosenberg, M.; Gwynn, M. N., Identification, evolution, and essentiality of the mevalonate pathway for isopentenyl diphosphate biosynthesis in gram-positive cocci. *Journal of Bacteriology* **2000,** *182* (15), 4319-4327.
213. Marrakchi, H.; Patel, D.; Rock, C., Regulation of fatty acid biosynthesis and degradation in Escherichia coli. *Faseb J* **2001,** *15* (4), A192-A192.
214. Mazein, A.; Watterson, S.; Hsieh, W. Y.; Griffiths, W. J.; Ghazal, P., A comprehensive machine-readable view of the mammalian cholesterol biosynthesis pathway. *Biochem Pharmacol* **2013,** *86* (1), 56-66.
215. Hampton, R.; DimsterDenk, D.; Rine, J., The biology of HMG-CoA reductase: The pros of contra-regulation. *Trends Biochem Sci* **1996,** *21* (4), 140-145.
216. Istvan, E. S.; Palnitkar, M.; Buchanan, S. K.; Deisenhofer, J., Crystal structure of the catalytic portion of human HMG-CoA reductase: insights into regulation of activity and catalysis. *Embo J* **2000,** *19* (5), 819-830.
217. Gesto, D. S.; Cerqueira, N. M.; Ramos, M. J.; Fernandes, P. A., Discovery of new druggable sites in the anti-cholesterol target HMG-CoA reductase by computational alanine scanning mutagenesis. *Journal of molecular modeling* **2014,** *20* (4), 2178.
218. Tabernero, L.; Bochar, D. A.; Rodwell, V. W.; Stauffacher, C. V., Substrate-induced closure of the flap domain in the ternary complex structures provides insights into the mechanism of catalysis by 3-hydroxy-3-methylglutaryl-CoA reductase. *P Natl Acad Sci USA* **1999,** *96* (13), 7167-7171.
219. Rodwell, V. W.; Beach, M. J.; Bischoff, K. M.; Bochar, D. A.; Darnay, B. G.; Friesen, J. A.; Gill, J. F.; Hedl, M.; Jordan-Starck, T.; Kennelly, P. J.; Kim, D.; Wang, Y. L., 3-hydroxy-3-methylglutaryl-CoA reductase. *Method Enzymol* **2000,** *324*, 259-280.
220. Istvan, E. S.; Deisenhofer, J., The structure of the catalytic portion of human HMG-CoA reductase. *Bba-Mol Cell Biol L* **2000,** *1529* (1-3), 9-18.
221. Veloso, D.; Cleland, W. W.; Porter, J. W., Ph Properties and Chemical Mechanism of Action of 3-Hydroxy-3-Methylglutaryl Coenzyme-a Reductase. *Biochemistry* **1981,** *20* (4), 887-894.
222. Wang, Y. L.; Darnay, B. G.; Rodwell, V. W., Identification of the Principal Catalytically Important Acidic Residue of 3-Hydroxy-3-Methylglutaryl Coenzyme-a Reductase. *Journal of Biological Chemistry* **1990,** *265* (35), 21634-21641.
223. Darnay, B. G.; Wang, Y. L.; Rodwell, V. W., Identification of the Catalytically Important Histidine of 3-Hydroxy-3-Methylglutaryl-Coenzyme-a Reductase. *Journal of Biological Chemistry* **1992,** *267* (21), 15064-15070.
224. Bochar, D. A.; Tabernero, L.; Stauffacher, C. V.; Rodwell, V. W., Aminoethylcysteine can replace the function of the essential active site lysine of Pseudomonas mevalonii 3-hydroxy-3-methylglutaryl coenzyme A reductase. *Biochemistry* **1999,** *38* (28), 8879-8883.
225. Frimpong, K.; Rodwell, V. W., Catalysis by Syrian-Hamster 3-Hydroxy-3-Methylglutaryl-Coenzyme-a Reductase - Proposed Roles of Histidine-865, Glutamate-558, and Aspartate-766. *Journal of Biological Chemistry* **1994,** *269* (15), 11478-11483.
226. Haines, B. E.; Wiest, O.; Stauffacher, C. V., The Increasingly Complex Mechanism of HMG-CoA Reductase. *Accounts Chem Res* **2013,** *46* (11), 2416-2426.
227. Friesen, J. A.; Rodwell, V. W., The 3-hydroxy-3-methylglutaryl coenzyme-A (HMG-CoA) reductases. *Genome Biol* **2004,** *5* (11).

228. Chambers, C. M.; Ness, G. C., Dietary cholesterol regulates hepatic 3-hydroxy-3-methylglutaryl coenzyme A reductase gene expression in rats primarily at the level of translation. *Arch Biochem Biophys* **1998,** *354* (2), 317-322.
229. Ness, G. C., Physiological feedback regulation of cholesterol biosynthesis: Role of translational control of hepatic HMG-CoA reductase and possible involvement of oxylanosterols. *Bba-Mol Cell Biol L* **2015,** *1851* (5), 667-673.
230. Chambers, C. M.; Ness, G. C., Translational regulation of hepatic HMG-CoA reductase by dietary cholesterol. *Biochemical and biophysical research communications* **1997,** *232* (2), 278-81.
231. Keller, R. K.; Zhao, Z.; Chambers, C.; Ness, G. C., Farnesol Is Not the Nonsterol Regulator Mediating Degradation of HMG-CoA Reductase in Rat Liver. *Arch Biochem Biophys* **1996,** *328* (2), 324–330.
232. Haines, B. E.; Wiest, O.; Stauffacher, C. V., The Increasingly Complex Mechanism of HMG-CoA Reductase. *Accounts of chemical research* **2013**.
233. Ness, G. C.; Lopez, D.; Chambers, C. M.; Zhao, Z.; Beach, D. L.; Ko, S. S.; Trzaskos, J. M., Effects of 15-oxa-32-vinyl-lanost-8-ene-3 beta,32 diol on the expression of 3-hydroxy-3-methylglutaryl coenzyme A reductase and low density lipoprotein receptor in rat liver. *Arch Biochem Biophys* **1998,** *357* (2), 259-64.
234. Farmer, A. R.; Murray, C. K.; Mende, K.; Akers, K. S.; Zera, W. C.; Beckius, M. L.; Yun, H. C., Effect of HMG-CoA reductase inhibitors on antimicrobial susceptibilities for Gram-Negative rods. *J Basic Microb* **2013,** *53* (4), 336-339.
235. Nalin, D. R., Comment on: Unexpected antimicrobial effect of statins. *J Antimicrob Chemoth* **2008,** *61* (6), 1400-1400.
236. Jerwood, S.; Cohen, J., Unexpected antimicrobial effect of statins. *J Antimicrob Chemoth* **2008,** *61* (2), 362-364.
237. Hogenboom, S.; Tuyp, J. J. M.; Espeel, M.; Koster, J.; Wanders, R. J. A.; Waterham, H. R., Phosphomevalonate kinase is a cytosolic protein in humans. *J Lipid Res* **2004,** *45* (4), 697-705.
238. Hogenboom, S.; Tuyp, J. J. M.; Espeel, M.; Koster, J.; Wanders, R. J. A.; Waterham, H. R., Human mevalonate pyrophosphate decarboxylase is localized in the cytosol. *Molecular Genetics and Metabolism* **2004,** *81* (3), 216-224.
239. Hogenboom, S.; Tuyp, J. J. M.; Espeel, M.; Koster, J.; Wanders, R. J. A.; Waterham, H. R., Mevalonate kinase is a cytosolic enzyme in humans. *J Cell Sci* **2004,** *117* (4), 631-639.
240. Olivier, L. M.; Chambliss, K. L.; Gibson, K. M.; Krisans, S. K., Characterization of phosphomevalonate kinase: chromosomal localization, regulation, and subcellular targeting. *J Lipid Res* **1999,** *40* (4), 672-679.
241. Kovacs, W. J.; Olivier, L. M.; Krisans, S. K., Central role of peroxisomes in isoprenoid biosynthesis. *Prog Lipid Res* **2002,** *41* (5), 369-391.
242. Fu, Z. J.; Wang, M.; Potter, D.; Miziorko, H. M.; Kim, J. J. P., The structure of a binary complex between a mammalian mevalonate kinase and ATP - Insights into the reaction mechanism and human inherited disease. *Journal of Biological Chemistry* **2002,** *277* (20), 18134-18142.
243. Fu, Z.; Voynova, N. E.; Herdendorf, T. J.; Miziorko, H. M.; Kim, J. J. P., Biochemical and structural basis for feedback inhibition of mevalonate kinase and isoprenoid metabolism. *Biochemistry* **2008,** *47* (12), 3715-3724.
244. Amdur, B. H.; Rilling, H.; Bloch, K., The Enzymatic Conversion of Mevalonic Acid to Squalene. *J Am Chem Soc* **1957,** *79* (10), 2646-2647.
245. Thurnher, M.; Nussbaumer, O.; Gruenbacher, G., Novel Aspects of Mevalonate Pathway Inhibitors as Antitumor Agents. *Clin Cancer Res* **2012,** *18* (13), 3524-3531.
246. Potter, D.; Miziorko, H. M., Identification of catalytic residues in human mevalonate kinase. *Journal of Biological Chemistry* **1997,** *272* (41), 25449-25454.
247. Beytia, E.; Dorsey, J. K.; Marr, J.; Cleland, W. W.; Porter, J. W., Purification and Mechanism of Action of Hog Liver Mevalonic Kinase. *Journal of Biological Chemistry* **1970,** *245* (20), 5450-&.

248. Cho, Y. K.; Rios, S. E.; Kim, J. J. P.; Miziorko, H. M., Investigation of invariant serine/threonine residues in mevalonate kinase - Tests of the functional significance of a proposed substrate binding motif and a site implicated in human inherited disease. *Journal of Biological Chemistry* **2001,** *276* (16), 12573-12578.
249. Potter, D.; Wojnar, J. M.; Narasimhan, C.; Miziorko, H. M., Identification and functional characterization of an active-site lysine in mevalonate kinase. *Journal of Biological Chemistry* **1997,** *272* (9), 5741-5746.
250. Berger, S. A.; Evans, P. R., Site-Directed Mutagenesis Identifies Catalytic Residues in the Active-Site of Escherichia-Coli Phosphofructokinase. *Biochemistry* **1992,** *31* (38), 9237-9242.
251. Chang, Q.; Yan, X. X.; Gu, S. Y.; Liu, J. F.; Liang, D. C., Crystal structure of human phosphomavelonate kinase at 1.8 angstrom resolution. *Proteins* **2008,** *73* (1), 254-258.
252. Herdendorf, T. J.; Miziorko, H. M., Functional evaluation of conserved basic residues in human phosphomevalonate kinase. *Biochemistry* **2007,** *46* (42), 11780-11788.
253. Olson, A. L.; Yao, H. L.; Herdendorf, T. J.; Miziorko, H. M.; Hannongbua, S.; Saparpakorn, P.; Cai, S.; Sem, D. S., Substrate induced structural and dynamics changes in human phosphomevalonate kinase and implications for mechanism. *Proteins* **2009,** *75* (1), 127-138.
254. Leipe, D. D.; Koonin, E. V.; Aravind, L., Evolution and classification of P-loop kinases and related proteins. *J Mol Biol* **2003,** *333* (4), 781-815.
255. Koonin, E. V., A Superfamily of Atpases with Diverse Functions Containing Either Classical or Deviant Atp-Binding Motif (Vol 229, Pg 1165, 1993). *J Mol Biol* **1993,** *232* (3), 1013-1013.
256. Herdendorf, T. J.; Miziorko, H. M., Phosphomevalonate kinase: Functional investigation of the recombinant human enzyme. *Biochemistry* **2006,** *45* (10), 3235-3242.
257. Ramachandran, C. K.; Shah, S. N., Decarboxylation of Mevalonate Pyrophosphate Is One Rate-Limiting Step in Hepatic Cholesterol-Synthesis in Suckling and Weaned Rats. *Biochemical and biophysical research communications* **1976,** *69* (1), 42-47.
258. Sawamura, M.; Nara, Y.; Yamori, Y., Liver Mevalonate 5-Pyrophosphate Decarboxylase Is Responsible for Reduced Serum-Cholesterol in Stroke-Prone Spontaneously Hypertensive Rat. *Journal of Biological Chemistry* **1992,** *267* (9), 6051-6055.
259. Voynova, N. E.; Fu, Z. J.; Battaile, K. P.; Herdendorf, T. J.; Kim, J. J. P.; Miziorko, H. M., Human mevalonate diphosphate decarboxylase: Characterization, investigation of the mevalonate diphosphate binding site, and crystal structure. *Arch Biochem Biophys* **2008,** *480* (1), 58-67.
260. Barta, M. L.; Skaff, D. A.; McWhorter, W. J.; Herdendorf, T. J.; Miziorko, H. M.; Geisbrecht, B. V., Crystal Structures of Staphylococcus epidermidis Mevalonate Diphosphate Decarboxylase Bound to Inhibitory Analogs Reveal New Insight into Substrate Binding and Catalysis. *Journal of Biological Chemistry* **2011,** *286* (27), 23900-23910.
261. Krepkiy, D.; Miziorko, H. M., Identification of active site residues in mevalonate diphosphate decarboxylase: Implications for a family of phosphotransferases. *Protein Sci* **2004,** *13* (7), 1875-1881.
262. Krepkiy, D. V.; Miziorko, H. M., Investigation of the functional contributions of invariant serine residues in yeast mevalonate diphosphate decarboxylase. *Biochemistry* **2005,** *44* (7), 2671-2677.
263. Barta, M. L.; McWhorter, W. J.; Miziorko, H. M.; Geisbrecht, B. V., Structural basis for nucleotide binding and reaction catalysis in mevalonate diphosphate decarboxylase. *Biochemistry* **2012,** *51* (28), 5611-21.

264. Steinbacher, S.; Kaiser, J.; Gerhardt, S.; Eisenreich, W.; Huber, R.; Bacher, A.; Rohdich, F., Crystal structure of the type II isopentenyl diphosphate: Dimethylallyl diphosphate isomerase from Bacillus subtilis. *J Mol Biol* **2003,** *329* (5), 973-982.
265. Durbecq, V.; Sainz, G.; Oudjama, Y.; Clantin, B.; Bompard-Gilles, C.; Tricot, C.; Caillet, J.; Stalon, V.; Droogmans, L.; Villeret, V., Crystal structure of isopentenyl diphosphate : dimethylallyl diphosphate isomerase. *Embo Journal* **2001,** *20* (7), 1530-1537.
266. Wouters, J.; Oudjama, Y.; Ghosh, S.; Stalon, V.; Droogmans, L.; Oldfield, E., Structure and mechanism of action of isopentenylpyrophosphate-dimethylallylpyrophosphate isomerase. *J Am Chem Soc* **2003,** *125* (11), 3198-3199.
267. Reardon, J. E.; Abeles, R. H., Mechanism of Action of Isopentenyl Pyrophosphate Isomerase - Evidence for a Carbonium-Ion Intermediate. *Biochemistry* **1986,** *25* (19), 5609-5616.
268. Street, I. P.; Christensen, D. J.; Poulter, C. D., Hydrogen-Exchange during the Enzyme-Catalyzed Isomerization of Isopentenyl Diphosphate and Dimethylallyl Diphosphate. *J Am Chem Soc* **1990,** *112* (23), 8577-8578.
269. Chapman, M. A.; Lawrence, M. S.; Keats, J. J.; Cibulskis, K.; Sougnez, C.; Schinzel, A. C.; Harview, C. L.; Brunet, J. P.; Ahmann, G. J.; Adli, M.; Anderson, K. C.; Ardlie, K. G.; Auclair, D.; Baker, A.; Bergsagel, P. L.; Bernstein, B. E.; Drier, Y.; Fonseca, R.; Gabriel, S. B.; Hofmeister, C. C.; Jagannath, S.; Jakubowiak, A. J.; Krishnan, A.; Levy, J.; Liefeld, T.; Lonial, S.; Mahan, S.; Mfuko, B.; Monti, S.; Perkins, L. M.; Onofrio, R.; Pugh, T. J.; Rajkumar, S. V.; Ramos, A. H.; Siegel, D. S.; Sivachenko, A.; Stewart, A. K.; Trudel, S.; Vij, R.; Voet, D.; Winckler, W.; Zimmerman, T.; Carpten, J.; Trent, J.; Hahn, W. C.; Garraway, L. A.; Meyerson, M.; Lander, E. S.; Getz, G.; Golub, T. R., Initial genome sequencing and analysis of multiple myeloma. *Nature* **2011,** *471* (7339), 467-472.
270. Coleman, R. E., Clinical features of metastatic bone disease and risk of skeletal morbidity. *Clin Cancer Res* **2006,** *12* (20), 6243S-6249S.
271. Fournier, P. G.; Stresing, V.; Ebetino, F. H.; Clezardin, P., How Do Bisphosphonates Inhibit Bone Metastasis In Vivo? *Neoplasia* **2010,** *12* (7), 571-578.
272. Leung, C. Y.; Park, J.; De Schutter, J. W.; Sebag, M.; Berghuis, A. M.; Tsantrizos, Y. S., Thienopyrimidine Bisphosphonate (ThPBP) Inhibitors of the Human Farnesyl Pyrophosphate Synthase: Optimization and Characterization of the Mode of Inhibition. *J Med Chem* **2013,** *56* (20), 7939-7950.
273. Monkkonen, H.; Auriola, S.; Lehenkari, P.; Kellinsalmi, M.; Hassinen, I. E.; Vepsalainen, J.; Monkkonen, J., A new endogenous ATP analog (ApppI) inhibits the mitochondrial adenine nucleotide translocase (ANT) and is responsible for the apoptosis induced by nitrogen-containing bisphosphonates. *Brit J Pharmacol* **2006,** *147* (4), 437-445.
274. Mitrofan, L. M.; Pelkonen, J.; Monkkonen, J., The level of ATP analog and isopentenyl pyrophosphate correlates with zoledronic acid-induced apoptosis in cancer cells in vitro. *Bone* **2009,** *45* (6), 1153-1160.
275. Russell, R. G. G., Bisphosphonates: The first 40 years. *Bone* **2011,** *49* (1), 2-19.
276. Szkopinska, A.; Plochocka, D., Farnesyl diphosphate synthase; regulation of product specificity. *Acta Biochim Pol* **2005,** *52* (1), 45-55.
277. Krisans, S. K.; Ericsson, J.; Edwards, P. A.; Keller, G. A., Farnesyl-Diphosphate Synthase Is Localized in Peroxisomes. *Journal of Biological Chemistry* **1994,** *269* (19), 14165-14169.
278. Song, L. S.; Poulter, C. D., Yeast Farnesyl-Diphosphate Synthase - Site-Directed Mutagenesis of Residues in Highly Conserved Prenyltransferase Domain-I and Domain-Ii. *P Natl Acad Sci USA* **1994,** *91* (8), 3044-3048.
279. Poulter, C. D.; Mash, E. A.; Argyle, J. C.; Muscio, O. J.; Rilling, H. C., Farnesyl Pyrophosphate Synthetase - Mechanistic Studies of the 1'-4 Coupling Reaction in the Terpene Biosynthetic-Pathway. *J Am Chem Soc* **1979,** *101* (22), 6761-6763.

280. Tarshis, L. C.; Proteau, P. J.; Kellogg, B. A.; Sacchettini, J. C.; Poulter, C. D., Regulation of product chain length by isoprenyl diphosphate synthases. *Proc Natl Acad Sci U S A* **1996,** *93* (26), 15018-23.
281. Hosfield, D. J.; Zhang, Y. M.; Dougan, D. R.; Broun, A.; Tari, L. W.; Swanson, R. V.; Finn, J., Structural basis for bisphosphonate-mediated inhibition of isoprenoid biosynthesis. *Journal of Biological Chemistry* **2004,** *279* (10), 8526-8529.
282. Sanchez, V. M.; Crespo, A.; Gutkind, J. S.; Turjanski, A. G., Investigation of the catalytic mechanism of farnesyl pyrophosphate synthase by computer simulation. *The journal of physical chemistry. B* **2006,** *110* (36), 18052-7.
283. Pandit, J.; Danley, D. E.; Schulte, G. K.; Mazzalupo, S.; Pauly, T. A.; Hayward, C. M.; Hamanaka, E. S.; Thompson, J. F.; Harwood, H. J., Crystal structure of human squalene synthase - A key enzyme in cholesterol biosynthesis. *Journal of Biological Chemistry* **2000,** *275* (39), 30610-30617.
284. Poulter, C. D., Biosynthesis of Non-Head-to-Tail Terpenes - Formation of 1'-1 and 1'-3 Linkages. *Accounts Chem Res* **1990,** *23* (3), 70-77.
285. Sandifer, R. M.; Thompson, M. D.; Gaughan, R. G.; Poulter, C. D., Squalene Synthetase - Inhibition by an Ammonium Analog of a Carbocationic Intermediate in the Conversion of Presqualene Pyrophosphate to Squalene. *J Am Chem Soc* **1982,** *104* (25), 7376-7378.
286. Gu, P.; Ishii, Y.; Spencer, T. A.; Shechter, I., Function-structure studies and identification of three enzyme domains involved in the catalytic activity in rat hepatic squalene synthase. (vol 273, pg 12515, 1998). *Journal of Biological Chemistry* **1998,** *273* (27), 17296-17296.
287. Liu, C. I.; Liu, G. Y.; Song, Y. C.; Yin, F. L.; Hensler, M. E.; Jeng, W. Y.; Nizet, V.; Wang, A. H. J.; Oldfield, E., A cholesterol biosynthesis inhibitor blocks Staphylococcus aureus virulence. *Science* **2008,** *319* (5868), 1391-1394.
288. Lin, F. Y.; Liu, C. I.; Liu, Y. L.; Zhang, Y. H.; Wang, K.; Jeng, W. Y.; Ko, T. P.; Cao, R.; Wang, A. H. J.; Oldfield, E., Mechanism of action and inhibition of dehydrosqualene synthase. *P Natl Acad Sci USA* **2010,** *107* (50), 21337-21342.
289. Do, R.; Kiss, R. S.; Gaudet, D.; Engert, J. C., Squalene synthase: a critical enzyme in the cholesterol biosynthesis pathway. *Clin Genet* **2009,** *75* (1), 19-29.
290. Ness, G. C.; Zhao, Z. H.; Keller, R. K., Effect of Squalene Synthase Inhibition on the Expression of Hepatic Cholesterol Biosynthetic-Enzymes, Ldl Receptor, and Cholesterol 7-Alpha Hydroxylase. *Arch Biochem Biophys* **1994,** *311* (2), 277-285.
291. Bergstrom, J. D.; Kurtz, M. M.; Rew, D. J.; Amend, A. M.; Karkas, J. D.; Bostedor, R. G.; Bansal, V. S.; Dufresne, C.; Vanmiddlesworth, F. L.; Hensens, O. D.; Liesch, J. M.; Zink, D. L.; Wilson, K. E.; Onishi, J.; Milligan, J. A.; Bills, G.; Kaplan, L.; Omstead, M. N.; Jenkins, R. G.; Huang, L.; Meinz, M. S.; Quinn, L.; Burg, R. W.; Kong, Y. L.; Mochales, S.; Mojena, M.; Martin, I.; Pelaez, F.; Diez, M. T.; Alberts, A. W., Zaragozic Acids - a Family of Fungal Metabolites That Are Picomolar Competitive Inhibitors of Squalene Synthase. *P Natl Acad Sci USA* **1993,** *90* (1), 80-84.
292. Stein, E. A.; Bays, H.; O'Brien, D.; Pedicano, J.; Piper, E.; Spezzi, A., Lapaquistat acetate: development of a squalene synthase inhibitor for the treatment of hypercholesterolemia. *Circulation* **2011,** *123* (18), 1974-85.
293. Ichikawa, M.; Ohtsuka, M.; Ohki, H.; Ota, M.; Haginoya, N.; Itoh, M.; Shibata, Y.; Ishigai, Y.; Terayama, K.; Kanda, A.; Sugita, K., Discovery of DF-461, a Potent Squalene Synthase Inhibitor. *ACS medicinal chemistry letters* **2013,** *4* (10), 932-6.
294. Stein, E. A.; Bays, H.; O'Brien, D.; Pedicano, J.; Piper, E.; Spezzi, A., Lapaquistat Acetate Development of a Squalene Synthase Inhibitor for the Treatment of Hypercholesterolemia. *Circulation* **2011,** *123* (18), 1974-1985.
295. Urbina, J. A.; Concepcion, J. L.; Rangel, S.; Visbal, G.; Lira, R., Squalene synthase as a chemotherapeutic target in Trypanosoma cruzi and Leishmania mexicana. *Mol Biochem Parasit* **2002,** *125* (1-2), 35-45.
296. Urbina, J. A., Specific treatment of Chagas disease: current status and new developments. *Curr Opin Infect Dis* **2001,** *14* (6), 733-741.

297. Urbina, J. A., Lipid biosynthesis pathways as chemotherapeutic targets in kinetoplastid parasites. *Parasitology* **1997,** *114*, S91-S99.
298. Urbina, J. A.; Concepcion, J. L.; Caldera, A.; Payares, G.; Sanoja, C.; Otomo, T.; Hiyoshi, H., In vitro and in vivo activities of E5700 and ER-119884, two novel orally active squalene synthase inhibitors, against Trypanosoma cruzi. *Antimicrob Agents Ch* **2004,** *48* (7), 2379-2387.
299. Urbina, J. A., Ergosterol biosynthesis and drug development for Chagas disease. *Mem I Oswaldo Cruz* **2009,** *104*, 311-318.
300. McTaggart, F.; Brown, G. R.; Davidson, R. G.; Freeman, S.; Holdgate, G. A.; Mallion, K. B.; Mirrlees, D. J.; Smith, G. J.; Ward, W. H. J., Inhibition of squalene synthase of rat liver by novel 3' substituted quinuclidines. *Biochem Pharmacol* **1996,** *51* (11), 1477-1487.
301. Goldstein, J. L.; Brown, M. S., Molecular medicine - The cholesterol quartet. *Science* **2001,** *292* (5520), 1310-1312.
302. Ono, T.; Bloch, K., Solubilization and Partial Characterization of Rat-Liver Squalene Epoxidase. *Journal of Biological Chemistry* **1975,** *250* (4), 1571-1579.
303. Gotteland, J. P.; Junquero, D.; Oms, P.; Delhon, A.; Halazy, S., Comparative study of new and known thienyl derivatives of ENE-YNE benzylamine as mammalian squalene epoxidase inhibitors. *Med Chem Res* **1996,** *6* (5), 333-342.
304. Astruc, M.; Tabacik, C.; Descomps, B.; Crastesdepaulet, A., Squalene Epoxidase and Oxidosqualene Lanosterol-Cyclase Activities in Cholesterogenic and Non-Cholesterogenic Tissues. *Biochimica Et Biophysica Acta* **1977,** *487* (1), 204-211.
305. Nagumo, A.; Kamei, T.; Sakakibara, J.; Ono, T., Purification and Characterization of Recombinant Squalene Epoxidase. *J Lipid Res* **1995,** *36* (7), 1489-1497.
306. Abe, I.; Seki, T.; Noguchi, H., Potent and selective inhibition of squalene epoxidase by synthetic galloyl esters. *Biochemical and biophysical research communications* **2000,** *270* (1), 137-140.
307. Laden, B. P.; Porter, T. D., Inhibition of human squalene monooxygenase by tellurium compounds: evidence of interaction with vicinal sulfhydryls. *J Lipid Res* **2001,** *42* (2), 235-240.
308. Chugh, A.; Ray, A.; Gupta, J. B., Squalene epoxidase as hypocholesterolemic drug target revisited. *Prog Lipid Res* **2003,** *42* (1), 37-50.
309. Belter, A.; Skupinska, M.; Giel-Pietraszuk, M.; Grabarkiewicz, T.; Rychlewski, L.; Barciszewski, J., Squalene monooxygenase - a target for hypercholesterolemic therapy. *Biol Chem* **2011,** *392* (12), 1053-1075.
310. Ryder, N. S., Terbinafine - Mode of Action and Properties of the Squalene Epoxidase Inhibition. *Brit J Dermatol* **1992,** *126*, 2-7.
311. Horie, M.; Tsuchiya, Y.; Hayashi, M.; Iida, Y.; Iwasawa, Y.; Nagata, Y.; Sawasaki, Y.; Fukuzumi, H.; Kitani, K.; Kamei, T., Nb-598 - a Potent Competitive Inhibitor of Squalene Epoxidase. *Journal of Biological Chemistry* **1990,** *265* (30), 18075-18078.
312. Horie, M.; Sawasaki, Y.; Fukuzumi, H.; Watanabe, K.; Iizuka, Y.; Tsuchiya, Y.; Kamei, T., Hypolipidemic Effects of Nb-598 in Dogs. *Atherosclerosis* **1991,** *88* (2-3), 183-192.
313. Thoma, R.; Schulz-Gasch, T.; D'Arcy, B.; Benz, J.; Aebi, J.; Dehmlow, H.; Hennig, M.; Stihle, M.; Ruf, A., Insight into steroid scaffold formation from the structure of human oxidosqualene cyclase. *Nature* **2004,** *432* (7013), 118-122.
314. Nakano, C.; Motegi, A.; Sato, T.; Onodera, M.; Hoshino, T., Sterol biosynthesis by a prokaryote: First in vitro identification of the genes encoding squalene epoxidase and lanosterol synthase from Methylococcus capsulatus. *Biosci Biotech Bioch* **2007,** *71* (10), 2543-2550.
315. Ruf, A.; Muller, F.; D'Arcy, B.; Stihle, M.; Kusznir, E.; Handschin, C.; Morand, O. H.; Thoma, R., The monotopic membrane protein human oxidosqualene cyclase is active as monomer. *Biochemical and biophysical research communications* **2004,** *315* (2), 247-254.

316. Abe, I.; Rohmer, M.; Prestwich, G. D., Enzymatic Cyclization of Squalene and Oxidosqualene to Sterols and Triterpenes. *Chem Rev* **1993,** *93* (6), 2189-2206.
317. Hess, B. A., Concomitant C-ring expansion and D-ring formation in lanosterol biosynthesis from squalene without violation of Markovnikov's rule. *J Am Chem Soc* **2002,** *124* (35), 10286-10287.
318. van Tamelen, E. E.; Willett, J. D.; Clayton, R. B., On the mechanism of lanosterol biosynthesis from squalene 2,3-oxide. *J Am Chem Soc* **1967,** *89* (13), 3371-3.
319. Vantamel.Ee; Willet, J.; Schwartz, M.; Nadeau, R., Nonenzymic Laboratory Cyclization of Squalene 2,3-Oxide. *J Am Chem Soc* **1966,** *88* (24), 5937-&.
320. Vantamel.Ee; Lees, R. G.; Grieder, A., Cyclization of a Terpenoid Diene with Preformed a-B-D Rings and Its Significance for Mechanism of Terpenoid Terminal Epoxide Cyclizations. *J Am Chem Soc* **1974,** *96* (7), 2255-2256.
321. Vantamelen, E. E.; James, D. R., Overall Mechanism of Terpenoid Terminal Epoxide Polycyclizations. *J Am Chem Soc* **1977,** *99* (3), 950-952.
322. Vantamelen, E. E.; Hopla, R. E., Generation of the Onocerin System by Lanosterol 2,3-Oxidosqualene Cyclase - Implications for the Cyclization Process. *J Am Chem Soc* **1979,** *101* (20), 6112-6114.
323. Vantamelen, E. E., Bioorganic Characterization and Mechanism of the 2,3-Oxidosqualene-] Lanosterol Conversion. *J Am Chem Soc* **1982,** *104* (23), 6480-6481.
324. Wendt, K. U.; Schulz, G. E.; Corey, E. J.; Liu, D. R., Enzyme mechanisms for polycyclic triterpene formation. *Angew Chem Int Edit* **2000,** *39* (16), 2812-+.
325. Tian, B. X.; Eriksson, L. A., Catalytic Mechanism and Product Specificity of Oxidosqualene-Lanosterol Cyclase: A QM/MM Study. *Journal of Physical Chemistry B* **2012,** *116* (47), 13857-13862.
326. Cory, E. J.; Russey, W. E.; Ortiz de Montellano, P. R., 2,3-oxidosqualene, an intermediate in the biological synthesis of sterols from squalene. *J Am Chem Soc* **1966,** *88* (20), 4750-1.
327. Jenson, C.; Jorgensen, W. L., Computational investigations of carbenium ion reactions relevant to sterol biosynthesis. *J Am Chem Soc* **1997,** *119* (44), 10846-10854.
328. Wu, T. K.; Wang, T. T.; Chang, C. H.; Liu, Y. T.; Shie, W. S., Importance of Saccharomyces cerevisiae oxidosqualene-lanosterol cyclase tyrosine 707 residue for chair-boat bicyclic ring formation and deprotonation reactions. *Organic letters* **2008,** *10* (21), 4959-62.
329. Wendt, K. U.; Lenhart, A.; Schulz, G. E., The structure of the membrane protein squalene-hopene cyclase at 2.0 A resolution. *J Mol Biol* **1999,** *286* (1), 175-87.
330. Cerqueira, N. M. F. S. A.; Fernandes, P. A.; Ramos, M. J., Computational Mechanistic Studies Addressed to the Transimination Reaction Present in All Pyridoxal 5 '-Phosphate-Requiring Enzymes. *J Chem Theory Comput* **2011,** *7* (5), 1356-1368.
331. Gesto, D. S.; Cerqueira, N. M.; Fernandes, P. A.; Ramos, M. J., Unraveling the enigmatic mechanism of L-asparaginase II with QM/QM calculations. *J Am Chem Soc* **2013,** *135* (19), 7146-58.
332. Ngo, H. P.; Cerqueira, N. M.; Kim, J. K.; Hong, M. K.; Fernandes, P. A.; Ramos, M. J.; Kang, L. W., PLP undergoes conformational changes during the course of an enzymatic reaction. *Acta crystallographica. Section D, Biological crystallography* **2014,** *70* (Pt 2), 596-606.
333. Oliveira, E. F.; Cerqueira, N. M.; Fernandes, P. A.; Ramos, M. J., Mechanism of formation of the internal aldimine in pyridoxal 5'-phosphate-dependent enzymes. *J Am Chem Soc* **2011,** *133* (39), 15496-505.
334. Ramos, M. J.; Fernandes, P. A., Computational enzymatic catalysis. *Acc Chem Res* **2008,** *41* (6), 689-98.
335. Sousa, S. F.; Fernandes, P. A.; Ramos, M. J., Computational enzymatic catalysis--clarifying enzymatic mechanisms with the help of computers. *Physical chemistry chemical physics : PCCP* **2012,** *14* (36), 12431-41.

336. Wilcox, C.; Turner, J.; Green, J., Systematic review: the management of chronic diarrhoea due to bile acid malabsorption. *Alimentary Pharmacology & Therapeutics* **2014,** *39* (9), 923-939.
337. Cannon, C. P.; Blazing, M. A.; Giugliano, R. P.; McCagg, A.; White, J. A.; Theroux, P.; Darius, H.; Lewis, B. S.; Ophuis, T. O.; Jukema, J. W.; De Ferrari, G. M.; Ruzyllo, W.; De Lucca, P.; Im, K.; Bohula, E. A.; Reist, C.; Wiviott, S. D.; Tershakovec, A. M.; Musliner, T. A.; Braunwald, E.; Califf, R. M.; Investigators, I.-I., Ezetimibe Added to Statin Therapy after Acute Coronary Syndromes. *The New England journal of medicine* **2015,** *372* (25), 2387-97.
338. Ness, G. C.; Holland, R. C.; Lopez, D., Selective compensatory induction of hepatic HMG-CoA reductase in response to inhibition of cholesterol absorption. *Experimental biology and medicine* **2006,** *231* (5), 559-65.
339. Li, C.; Lin, L.; Zhang, W.; Zhou, L.; Wang, H.; Luo, X.; Luo, H.; Cai, Y.; Zeng, C., Efficiency and safety of proprotein convertase subtilisin/kexin 9 monoclonal antibody on hypercholesterolemia: a meta-analysis of 20 randomized controlled trials. *Journal of the American Heart Association* **2015,** *4* (6), e001937.
340. Navarese, E. P.; Kolodziejczak, M.; Schulze, V.; Gurbel, P. A.; Tantry, U.; Lin, Y.; Brockmeyer, M.; Kandzari, D. E.; Kubica, J. M.; D'Agostino, R. B., Sr.; Kubica, J.; Volpe, M.; Agewall, S.; Kereiakes, D. J.; Kelm, M., Effects of Proprotein Convertase Subtilisin/Kexin Type 9 Antibodies in Adults With Hypercholesterolemia: A Systematic Review and Meta-analysis. *Annals of internal medicine* **2015,** *163* (1), 40-51.
341. Gesto, D. S.; Cerqueira, N. M. F. S. A.; Fernandes, P. A.; Ramos, M. J., Unraveling the Enigmatic Mechanism of L-Asparaginase II with QM/QM Calculations. *J Am Chem Soc* **2013,** *135* (19), 7146-7158.
342. Dinndorf, P. A.; Gootenberg, J.; Cohen, M. H.; Keegan, P.; Pazdur, R., FDA drug approval summary: Pegaspargase (Oncaspar (R)) for the first-line treatment of children with acute lymphoblastic leukemia (ALL). *Oncologist* **2007,** *12* (8), 991-998.
343. Verma, N.; Kumar, K.; Kaur, G.; Anand, S., L-asparaginase: A promising chemotherapeutic agent. *Crit Rev Biotechnol* **2007,** *27* (1), 45-62.
344. Hill, J. M.; Roberts, J.; Loeb, E.; Khan, A.; MacLellan, A.; Hill, R. W., L-asparaginase therapy for leukemia and other malignant neoplasms. *JAMA: the journal of the American Medical Association* **1967,** *202* (9), 882-888.
345. Beard, M.; Crowther, D.; Galton, D.; Guyer, R.; Fairley, G. H.; Kay, H.; Knapton, P.; Malpas, J.; Scott, R. B., L-asparaginase in treatment of acute leukaemia and lymphosarcoma. *Brit Med J* **1970,** *1* (5690), 191.
346. Kobrinsky, N. L.; Sposto, R.; Shah, N. R.; Anderson, J. R.; DeLaat, C.; Morse, M.; Warkentin, P.; Gilchrist, G. S.; Cohen, M. D.; Shina, D., Outcomes of treatment of children and adolescents with recurrent non-Hodgkin's lymphoma and Hodgkin's disease with dexamethasone, etoposide, cisplatin, cytarabine, and l-asparaginase, maintenance chemotherapy, and transplantation: Children's Cancer Group Study CCG-5912. *Journal of clinical oncology* **2001,** *19* (9), 2390-2396.
347. Broome, J., Studies on the mechanism of tumor inhibition by L-asparaginase. *The Journal of experimental medicine* **1968,** *127* (6), 1055.
348. Ho, P. P. K.; Milikin, E. B.; Bobbitt, J. L.; Grinnan, E. L.; Burck, P. J.; Frank, B. H.; Boeck, L. V. D.; Squires, R. W., Crystalline L-asparaginase from Escherichia coli B. *Journal of Biological Chemistry* **1970,** *245* (14), 3708-3715.
349. Ammon, H. L.; Weber, I. T.; Wlodawer, A.; Harrison, R. W.; Gilliland, G. L.; Murphy, K. C.; Sjolin, L.; Roberts, J., Preliminary crystal structure of Acinetobacter glutaminasificans glutaminase-asparaginase. *J Biol Chem* **1988,** *263* (1), 150-6.
350. Lubkowski, J.; Wlodawer, A.; Housset, D.; Weber, I. T.; Ammon, H. L.; Murphy, K. C.; Swain, A. L., Refined crystal structure of Acinetobacter glutaminasificans glutaminase-asparaginase. *Acta crystallographica. Section D, Biological crystallography* **1994,** *50* (Pt 6), 826-32.

351. Lubkowski, J.; Palm, G. J.; Gilliland, G. L.; Derst, C.; Rohm, K. H.; Wlodawer, A., Crystal structure and amino acid sequence of Wolinella succinogenes L-asparaginase. *European Journal of Biochemistry* **1996,** *241* (1), 201-7.
352. Swain, A. L.; Jaskolski, M.; Housset, D.; Rao, J. K. M.; Wlodawer, A., Crystal-Structure of Escherichia-Coli L-Asparaginase, an Enzyme Used in Cancer-Therapy. *P Natl Acad Sci USA* **1993,** *90* (4), 1474-1478.
353. Miller, M.; Rao, J. K. M.; Wlodawer, A.; Gribskov, M. R., A Left-Handed Crossover Involved in Amidohydrolase Catalysis - Crystal-Structure of Erwinia-Chrysanthemi L-Asparaginase with Bound L-Aspartate. *Febs Letters* **1993,** *328* (3), 275-279.
354. Aghaiypour, K.; Wlodawer, A.; Lubkowski, J., Structural basis for the activity and substrate specificity of Erwinia chrysanthemi L-asparaginase. *Biochemistry* **2001,** *40* (19), 5655-5664.
355. Ortlund, E.; Lacount, M. W.; Lewinski, K.; Lebioda, L., Reactions of Pseudomonas 7A glutaminase-asparaginase with diazo analogues of glutamine and asparagine result in unexpected covalent inhibitions and suggests an unusual catalytic triad Thr-Tyr-Glu. *Biochemistry* **2000,** *39* (6), 1199-204.
356. Harms, E.; Wehner, A.; Aung, H. P.; Rohm, K. H., A Catalytic Role for Threonine-12 of Escherichia-Coli Asparaginase-Ii as Established by Site-Directed Mutagenesis. *Febs Letters* **1991,** *285* (1), 55-58.
357. Derst, C.; Henseling, J.; Rohm, K. H., Probing the Role of Threonine and Serine Residues of Escherichia-Coli Asparaginase-Ii by Site-Specific Mutagenesis. *Protein Eng* **1992,** *5* (8), 785-789.
358. Palm, G. J.; Lubkowski, J.; Derst, C.; Schleper, S.; Rohm, K. H.; Wlodawer, A., A covalently bound catalytic intermediate in Escherichia coli asparaginase: Crystal structure of a Thr-89-val mutant. *Febs Letters* **1996,** *390* (2), 211-216.
359. Peterson, R. G.; Richards, F. F.; Handschumacher, R. E., Structure of Peptide from Active-Site Region of Escherichia-Coli L-Asparaginase. *Journal of Biological Chemistry* **1977,** *252* (6), 2072-2076.
360. Derst, C.; Wehner, A.; Specht, V.; Rohm, K. H., States and Functions of Tyrosine Residues in Escherichia-Coli Asparaginase-Ii. *European Journal of Biochemistry* **1994,** *224* (2), 533-540.
361. Morokuma, K.; Froese, R. D.; Dapprich, S.; Komaromi, I.; Khoroshun, D.; Byun, S.; Musaev, D. G.; Emerson, C. L., The ONIOM (our own integrated N-layered molecular orbital and molecular mechanics) method, and its applications to calculations of large molecular systems. *Abstr Pap Am Chem S* **1998,** *215*, U218-U218.
362. Becke, A. D., A New Mixing of Hartree-Fock and Local Density-Functional Theories. *J Chem Phys* **1993,** *98* (2), 1372-1377.
363. Lee, C. T.; Yang, W. T.; Parr, R. G., Development of the Colle-Salvetti Correlation-Energy Formula into a Functional of the Electron-Density. *Phys Rev B* **1988,** *37* (2), 785-789.
364. Vosko, S. H.; Wilk, L.; Nusair, M., Accurate Spin-Dependent Electron Liquid Correlation Energies for Local Spin-Density Calculations - a Critical Analysis. *Can J Phys* **1980,** *58* (8), 1200-1211.
365. Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J., Ab-Initio Calculation of Vibrational Absorption and Circular-Dichroism Spectra Using Density-Functional Force-Fields. *J Phys Chem-Us* **1994,** *98* (45), 11623-11627.
366. Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, J., J. A.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, N. J.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.;

Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, Ö.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. *Gaussian, Inc.*, Wallingford CT: 2009.
367. Cerqueira, N.; Fernandes, P.; Ramos, M., Computational Mechanistic Studies Addressed to the Transimination Reaction Present in All Pyridoxal 5′-Phosphate-Requiring Enzymes. *J Chem Theory Comput* **2011**.
368. Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P., The Development and Use of Quantum-Mechanical Molecular-Models .76. Am1 - a New General-Purpose Quantum-Mechanical Molecular-Model. *J Am Chem Soc* **1985,** *107* (13), 3902-3909.
369. Zhao, Y.; Truhlar, D. G., The M06 suite of density functionals for main group thermochemistry, thermochemical kinetics, noncovalent interactions, excited states, and transition elements: two new functionals and systematic testing of four M06-class functionals and 12 other functionals. *Theor Chem Acc* **2008,** *120* (1-3), 215-241.
370. Tomasi, J.; Mennucci, B.; Cammi, R., Quantum mechanical continuum solvation models. *Chem Rev* **2005,** *105* (8), 2999-3093.
371. Cerqueira, N. M. F. S. A.; Fernandes, P. A.; Eriksson, L. A.; Ramos, M. J., Dehydration of ribonucleotides catalyzed by ribonucleotide reductase: The role of the enzyme. *Biophys J* **2006,** *90* (6), 2109-2119.
372. Oliveira, E. F.; Cerqueira, N. M. F. S. A.; Fernandes, P. A.; Ramos, M. J., Mechanism of Formation of the Internal Aldimine in Pyridoxal 5 '-Phosphate-Dependent Enzymes. *J Am Chem Soc* **2011,** *133* (39), 15496-15505.
373. Aghaiypour, K.; Wlodawer, A.; Lubkowski, J., Do bacterial L-asparaginases utilize a catalytic triad Thr-Tyr-Glu? *Bba-Protein Struct M* **2001,** *1550* (2), 117-128.
374. Aung, H. P.; Bocola, M.; Schleper, S.; Rohm, K. H., Dynamics of a mobile loop at the active site of Escherichia coli asparaginase. *Bba-Protein Struct M* **2000,** *1481* (2), 349-359.
375. Gesto, D. S.; Cerqueira, N. M. F. S. A.; Ramos, M. J.; Fernandes, P. A., Discovery of new druggable sites in the anti-cholesterol target HMG-CoA reductase by computational alanine scanning mutagenesis. *Journal of molecular modeling* **2014,** *20* (4).
376. Black, D. M., Therapeutic targets in cardiovascular disease: A case for high-density lipoprotein cholesterol. *American Journal of Cardiology* **2003,** *91* (7A), 40E-43E.
377. Istvan, E., Statin inhibition of HMG-CoA reductase: a 3-dimensional view. *Atherosclerosis Supp* **2003,** *4* (1), 3-8.
378. Hermann, M.; Bogsrud, M. P.; Molden, E.; Asberg, A.; Mohebi, B. U.; Ose, L.; Retterstol, K., Exposure of atorvastatin is unchanged but lactone and acid metabolites are increased several-fold in patients with atorvastatin-induced myopathy. *Clin Pharmacol Ther* **2006,** *79* (6), 532-539.
379. Furberg, C. D.; Pitt, B., Withdrawal of cerivastatin from the world market. *Current controlled trials in cardiovascular medicine* **2001,** *2* (5), 205-207.
380. Massova, I.; Kollman, P. A., Computational Alanine Scanning To Probe Protein−Protein Interactions: A Novel Approach To Evaluate Binding Free Energies. *J Am Chem Soc* **1999,** *121* (36), 8133-8143.
381. Kortemme, T.; Baker, D., A simple physical model for binding energy hot spots in protein-protein complexes. *P Natl Acad Sci USA* **2002,** *99* (22), 14116-14121.
382. Li, H.; Robertson, A. D.; Jensen, J. H., Very fast empirical prediction and rationalization of protein pK(a) values. *Proteins* **2005,** *61* (4), 704-721.
383. Olsson, M. H. M.; Sondergaard, C. R.; Rostkowski, M.; Jensen, J. H., PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical pK(a) Predictions. *J Chem Theory Comput* **2011,** *7* (2), 525-537.
384. Case, D. A.; Darden, T. A.; Cheatham; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Merz, K. M.; Pearlman, D. A.; Crowley, M.; Walker, R. C.; Zhang, W.; Wang, B.; Hayik, S.; Roitberg, A.; Seabra, G.; Wong, K. F.; Paesani, F.; Wu, X.; Brozell, S.; Tsui,

V.; Gohlke, H.; Yang, L.; Tan, C.; Mongan, J.; Hornak, V.; Cui, G.; Beroza, P.; Mathews, D. H.; Schafmeister, C.; Ross, W. S.; Kollman., P. A. *AMBER9*, San Francisco, 2006.
385. Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C., Numerical-Integration of Cartesian Equations of Motion of a System with Constraints - Molecular-Dynamics of N-Alkanes. *J Comput Phys* **1977,** *23* (3), 327-341.
386. Martins, S. A.; Perez, M. A. S.; Moreira, I. S.; Sousa, S. F.; Ramos, M. J.; Fernandes, P. A., Computational Alanine Scanning Mutagenesis: MM-PBSA vs TI. *J Chem Theory Comput* **2013,** *9* (3), 1311-1319.
387. Moreira, I. S.; Fernandes, P. A.; Ramos, M. J., Unraveling the importance of protein-protein interaction: Application of a computational alanine-scanning mutagenesis to the study of the IgG1 streptococcal protein G (C2 fragment) complex. *Journal of Physical Chemistry B* **2006,** *110* (22), 10962-10969.
388. Moreira, I. S.; Fernandes, P. A.; Ramos, M. J., Hot spot occlusion from bulk water: A comprehensive study of the complex between the lysozyme HEL and the antibody FVD1.3. *Journal of Physical Chemistry B* **2007,** *111* (10), 2697-2706.
389. Moreira, I. S.; Fernandes, P. A.; Ramos, M. J., Protein-protein recognition: a computational mutagenesis study of the MDM2-P53 complex. *Theor Chem Acc* **2008,** *120* (4-6), 533-542.
390. Kollman, P. A.; Massova, I.; Reyes, C.; Kuhn, B.; Huo, S.; Chong, L.; Lee, M.; Lee, T.; Duan, Y.; Wang, W.; Donini, O.; Cieplak, P.; Srinivasan, J.; Case, D. A.; Cheatham, T. E., 3rd, Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models. *Acc Chem Res* **2000,** *33* (12), 889-97.
391. Moreira, I. S.; Fernandes, P. A.; Ramos, M. J., Computational alanine scanning mutagenesis - An improved methodological approach. *J Comput Chem* **2007,** *28* (3), 644-654.
392. Huo, S.; Massova, I.; Kollman, P. A., Computational alanine scanning of the 1:1 human growth hormone-receptor complex. *J Comput Chem* **2002,** *23* (1), 15-27.
393. Ribeiro, J. V.; Cerqueira, N. M. F. S. A.; Moreira, I. S.; Fernandes, P. A.; Ramos, M. J., CompASM: an Amber-VMD alanine scanning mutagenesis plug-in. *Theor Chem Acc* **2012,** *131*, 1271.
394. Humphrey, W.; Dalke, A.; Schulten, K., VMD - Visual Molecular Dynamics. *J. Molec. Graphics* **1996,** *14*, 33-38.

.
395. Ribeiro, J. V.; Cerqueira, N. M. F. S. A.; Fernandes, P. A.; Ramos, M. J., VOLAREA - a Bioinfomratic tool to calculate the surface area and the volume of molecular systems. *Chemical Biology* **2012**, (in press).
396. Moreira, I. S.; Fernandes, P. A.; Ramos, M. J., Hot spots-A review of the protein-protein interface determinant amino-acid residues. *Proteins* **2007,** *68* (4), 803-812.
397. Sousa, S. F.; Tamames, B.; Fernandes, P. A.; Ramos, M. J., Detailed Atomistic Analysis of the HIV-1 Protease Interface. *Journal of Physical Chemistry B* **2011,** *115* (21), 7045-7057.
398. Raper, H. S., The condensation of acetaldehyde and its relation to the biochemical synthesis of fatty acids. *J Chem Soc* **1907,** *91*, 1831-1838.
399. Klein, H. P.; Lipmann, F., The Relationship of Coenzyme a to Lipide Synthesis .2. Experiments with Rat Liver. *Journal of Biological Chemistry* **1953,** *203* (1), 101-108.
400. Brady, R. O., The Enzymatic Synthesis of Fatty Acids by Aldol Condensation. *P Natl Acad Sci USA* **1958,** *44* (10), 993-998.
401. Wakil, S. J., A Malonic Acid Derivative as an Intermediate in Fatty Acid Synthesis. *J Am Chem Soc* **1958,** *80* (23), 6465-6465.
402. Brindley, D. N.; Matsumur.S; Bloch, K., Mycobacterium Phlei Fatty Acid Synthetase - a Bacterial Multienzyme Complex. *Nature* **1969,** *224* (5220), 666-&.

403. Lomakin, I. B.; Xiong, Y.; Steitz, T. A., The crystal structure of yeast fatty acid synthase, a cellular machine with eight active sites working together. *Cell* **2007,** *129* (2), 319-332.
404. Reed, M. A. C.; Schweizer, M.; Szafranska, A. E.; Arthur, C.; Nicholson, T. P.; Cox, R. J.; Crosby, J.; Crump, M. P.; Simpson, T. J., The type I rat fatty acid synthase ACP shows structural homology and analogous biochemical properties to type II ACPs. *Org Biomol Chem* **2003,** *1* (3), 463-471.
405. Zhang, Y. M.; Wu, B. N.; Zheng, J.; Rock, C. O., Key residues responsible for acyl carrier protein and beta-ketoacyl-acyl carrier protein reductase (FabG) interaction. *Journal of Biological Chemistry* **2003,** *278* (52), 52935-52943.
406. Crump, M. P.; Crosby, J.; Dempsey, C. E.; Parkinson, J. A.; Murray, M.; Hopwood, D. A.; Simpson, T. J., Solution structure of the actinorhodin polyketide synthase acyl carrier protein from Streptomyces coelicolor A3(2). *Biochemistry* **1997,** *36* (20), 6000-6008.
407. Bunkoczi, G.; Pasta, S.; Joshi, A.; Wu, X. Q.; Kavanagh, K. L.; Smith, S.; Oppermann, U., Mechanism and substrate recognition of human holo ACP synthase. *Chem Biol* **2007,** *14* (11), 1243-1253.
408. Li, Q.; Khosla, C.; Puglisi, J. D.; Liu, C. W., Solution structure and backbone dynamics of the holo form of the frenolicin acyl carrier protein. *Biochemistry* **2003,** *42* (16), 4648-4657.
409. Smith, S.; Tsai, S. C., The type I fatty acid and polyketide synthases: a tale of two megasynthases. *Nat Prod Rep* **2007,** *24* (5), 1041-1072.
410. Stern, A.; Sedgwick, B.; Smith, S., The Free Coenzyme-a Requirement of Animal Fatty-Acid Synthetase - Participation in the Continuous Exchange of Acetyl and Malonyl Moieties between Co-Enzyme a Thioester and Enzyme. *Journal of Biological Chemistry* **1982,** *257* (2), 799-803.
411. Roujeinikova, A.; Simon, W. J.; Gilroy, J.; Rice, D. W.; Rafferty, J. B.; Slabas, A. R., Structural studies of fatty acyl-(acyl carrier protein) thioesters reveal a hydrophobic binding cavity that can expand to fit longer substrates. *J Mol Biol* **2007,** *365* (1), 135-145.
412. Witkowski, A.; Ghosal, A.; Joshi, A. K.; Witkowska, H. E.; Asturias, F. J.; Smith, S., Head-to-head coiled arrangement of the subunits of the animal fatty acid synthase. *Chem Biol* **2004,** *11* (12), 1667-1676.
413. Witkowski, A.; Joshi, A. K.; Smith, S., Mechanism of the beta-ketoacyl synthase reaction catalyzed by the animal fatty acid synthase. *Biochemistry* **2002,** *41* (35), 10877-10887.
414. Maier, T.; Leibundgut, M.; Ban, N., The crystal structure of a mammalian fatty acid synthase. *Science* **2008,** *321* (5894), 1315-1322.
415. Trott, O.; Olson, A. J., Software News and Update AutoDock Vina: Improving the Speed and Accuracy of Docking with a New Scoring Function, Efficient Optimization, and Multithreading. *J Comput Chem* **2010,** *31* (2), 455-461.
416. Field, M. J.; Bash, P. A.; Karplus, M., A Combined Quantum-Mechanical and Molecular Mechanical Potential for Molecular-Dynamics Simulations. *J Comput Chem* **1990,** *11* (6), 700-733.
417. Das, D.; Eurenius, K. P.; Billings, E. M.; Sherwood, P.; Chatfield, D. C.; Hodoscek, M.; Brooks, B. R., Optimization of quantum mechanical molecular mechanical partitioning schemes: Gaussian delocalization of molecular mechanical charges and the double link atom method. *J Chem Phys* **2002,** *117* (23), 10534-10547.
418. Sousa, S. F.; Ribeiro, A. J.; Neves, R. P.; Brás, N. F.; Cerqueira, N. M.; Fernandes, P. A.; Ramos, M. J., Application of quantum mechanics/molecular mechanics methods in the study of enzymatic reaction mechanisms. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2017,** *7* (2).